

# The American Economic Review

## ARTICLES

86

- C. L. SCHULTZE    Microeconomic Efficiency and Nominal Wage Stickiness  
S. BOWLES        The Production Process in a Competitive Economy  
G. J. BENSTON    The Validity of Profits-Structure with Particular Reference  
                         to the FTC's Line of Business Data  
P. EVANS         Do Large Deficits Produce High Interest Rates?  
P. HOWITT        Transaction Costs in the Theory of Unemployment  
R. H. FRANK      The Demand for Unobservable and Other Nonpositional Goods  
D. A. ASCHAUER   Fiscal Policy and Aggregate Demand  
C. L. BALLARD, J. B. SHOVEN, AND J. WHALLEY  
                         General Equilibrium Computations of the Marginal  
                         Welfare Costs of Taxes in the United States  
A. K. DIXIT AND A. S. KYLE  
                         The Use of Protection and Subsidies for  
                         Entry Promotion and Deterrence  
P. GOTTSCHALK AND S. DANZIGER  
                         A Framework for Evaluating the Effects of Economic Growth  
                         and Transfers on Poverty  
M. ESWARAN AND A. KOTWAL  
                         A Theory of Two-Tier Labor Markets in Agrarian Economies  
J. E. ANDERSON   The Relative Inefficiency of Quotas  
J. B. HAGENS AND R. R. RUSSELL  
                         Testing for the Effectiveness of Wage-Price Controls

SHORTER PAPERS: M. Adler; L. M. Nichols; M. K. Perry and R. H. Porter; C. Brown;  
R. Manning, J. R. Markusen, and J. McMillan; R. C. Fair; D. Léonard; P. Musgrove; G. Blazenko;  
J. A. K. Cave; B. P. Pesek; R. A. Heiner; D. Friedman; V. L. Smith; D. J. Richards;  
H. W. Chappell, Jr. and J. T. Addison, E. Briys and L. Eeckhoudt; J. D. Hey; E. Katz.

MARCH 1985

# THE AMERICAN ECONOMIC ASSOCIATION

Founded in 1885

## Officers

### President

CHARLES P. KINDLEBERGER

Massachusetts Institute of Technology

### President-Elect

ALICE M. MORTEN

The Brookings Institution

### Vice Presidents

ELIZABETH E. BAILEY

Carnegie-Mellon University

JOSEPH E. STIGLITZ

Princeton University

### Secretary

C. ELTON HINSHAW

Vanderbilt University

### Treasurer

RENDIGS FELS

Vanderbilt University

### Managing Editor of The American Economic Review

ROBERT W. CLOWER

University of California-Los Angeles

### Managing Editor of The Journal of Economic Literature

MOSES ABRAMOVITZ

Stanford University

## Executive Committee

### Elected Members of the Executive Committee

WILLIAM D. NORDHAUS

Yale University

A. MICHAEL SPENCE

Harvard University

VICTOR R. FUCHS

Stanford University

JANET L. NORWOOD

Bureau of Labor Statistics

ALAN S. BLINDER

Princeton University

DANIEL L. MCFADDEN

Massachusetts Institute of Technology

### EX OFFICIO Members

W. ARTHUR LEWIS

Princeton University

CHARLES L. SCHULTZE

The Brookings Institution

P 3125  
PC3425

●Published at George Banta Co., Inc., Menasha, Wisconsin. The publication number is ISSN 0002-8282.

●THE AMERICAN ECONOMIC REVIEW including four quarterly numbers, the *Proceedings* of the annual meetings, the *Directory*, and *Supplements*, is published by the American Economic Association and is sent to all members six times a year: March; May; June; September; semi-monthly, December.

Regular member dues for 1985, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature* are as follows:

\$35.00 if annual income is \$30,000 or less;

\$42.00 if annual income is above \$30,000, but no more than \$40,000;

\$49.00 if annual income is above \$40,000.

Nonmember subscriptions will be accepted only for both journals: Institutions (libraries, firms, etc.), \$100 a year; individuals, \$65.00. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.

In countries other than the United States, add \$11.00 to the annual rates above to cover extra postage.

●Correspondence relating to the *Directory*, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinstaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

●Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

●Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786.



# THE AMERICAN ECONOMIC REVIEW

391  
001

ROBERT W. CLOWER

Managing Editor

JOHN G. RILEY

Associate Editor

WILMA ST. JOHN

Production Editor

THERESA DE MARIA

Assistant Editor

## Board of Editors

GEORGE A. AKERLOF

CLIVE D. BULL

PATRICIA DANZON

MICHAEL R. DARBY

PHILIP E. GRAVES

MEIR KOHN

FREDERIC S. MISHKIN

SHERWIN ROSEN

RICHARD SCHMALENSEE

SUSAN WOODWARD

LESLIE YOUNG

•Beginning April 1, 1985, new manuscripts will be handled by Orley Ashenfelter of Princeton University, the incoming editor. New manuscripts should be addressed:

Orley Ashenfelter, Managing Editor, *American Economic Review*, 169 Nassau Street, Princeton, NJ 08540.

•Manuscripts should be submitted in triplicate and in acceptable form, and should be no longer than 50 pages of double-spaced typescript. A submission fee must accompany each manuscript: \$25 for members; \$50 for nonmembers. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

•No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

•Copyright © American Economic Association 1985. All rights reserved.

March 1985

VOLUME 75, NUMBER 1

## Articles

- Microeconomic Efficiency and Nominal Wage Stickiness *Charles L. Schultze* 1

- The Production Process in a Competitive Economy: Walrasian, Neo-Hobbesian, and Marxian Models *Samuel Bowles* 16

- The Validity of Profits-Structure Studies with Particular Reference to the FTC's Line of Business Data *George J. Benston* 37

- Do Large Deficits Produce High Interest Rates? *Paul Evans* 68

- Transaction Costs in the Theory of Unemployment *Peter Howitt* 88

- The Demand for Unobservable and Other Nonpositional Goods *Robert H. Frank* 101

- Fiscal Policy and Aggregate Demand *David Alan Aschauer* 117

- General Equilibrium Computations of the Marginal Welfare Costs of Taxes in the United States *Charles L. Ballard, John B. Shoven, and John Whalley* 128

- The Use of Protection and Subsidies for Entry Promotion and Deterrence *Avinash K. Dixit and Albert S. Kyle* 139

- A Framework for Evaluating the Effects of Economic Growth and Transfers on Poverty *Peter Gottschalk and Sheldon Danziger* 153

- A Theory of Two-Tier Labor Markets in Agrarian Economies *Mukesh Eswaran and Ashok Kotwal* 162

- The Relative Inefficiency of Quotas: The Cheese Case *James E. Anderson* 178

- Testing for the Effectiveness of Wage-Price Controls: An Application to the Carter Program *John B. Hagens and R. Robert Russell* 191

## Shorter Papers

Stardom and Talent	<i>Moshe Adler</i>	208
Advertising and Economic Welfare	<i>Len M. Nichols</i>	213
Oligopoly and the Incentive for Horizontal Merger	<i>Martin K. Perry and Robert H. Porter</i>	219
Military Enlistments: What Can We Learn from Geographic Variation?	<i>Charles Brown</i>	228
Paying for Public Inputs	<i>Richard Manning, James R. Markusen, and John McMillan</i>	235
Excess Labor and the Business Cycle	<i>Ray C. Fair</i>	239
Monopoly Unionism: Note	<i>Daniel Léonard</i>	246
Why Everything Takes 2.71828... Times as Long as Expected	<i>Philip Musgrove</i>	250
The Design of an Optimal Insurance Policy: Note	<i>George Blazenko</i>	253
A Further Comment on Preemptive Patenting and the Persistence of Monopoly	<i>Jonathan A. K. Cave</i>	256
A Curse on Several Houses	<i>Boris P. Pesek</i>	259
Experimental Economics:		
Comment	<i>Ronald A. Heiner</i>	260
Comment	<i>Daniel Friedman</i>	264
Reply	<i>Vernon L. Smith</i>	265
Relative Prices, Concentration, and Money Growth:		
Comment	<i>Daniel J. Richards</i>	273
Reply	<i>Henry W. Chappell, Jr. and John T. Addison</i>	281
Relative Risk Aversion in Comparative Statics:		
Comment	<i>Eric Briys and Louis Eeckhoudt</i>	284
Comment	<i>John D. Hey</i>	286
Reply	<i>Eliakim Katz</i>	288
Notes		288

P 3125

Number 86 of a series of photographs of past presidents of the Association

---



Charles L. Schultze

# Microeconomic Efficiency and Nominal Wage Stickiness<sup>†</sup>

By CHARLES L. SCHULTZE\*

...[T]he world may have its reasons for being non-Walrasian.

Robert Solow

The vast majority of our profession share a common view on most microeconomic policy issues. But we are widely split over macroeconomic theory and policy. Our consensus on micro issues arises from a shared model of how markets work in the long run. Our division on macro issues stems from a number of reasons, the emphasis on which has shifted over the years. In recent times the main disagreement has centered on how markets perform in the short run. In particular, we cannot agree on why nominal wages are sticky on the face of aggregate demand shocks. The traditional view argues that wages are structurally sticky. The new classical macroeconomists argue that wages are fundamentally flexible. But the rational expectations of economic agents, grounded on past experience with attempts at employment-supporting monetary policy, have produced the observed wage stickiness. Introduction of a changed policy regime, based on a stable growth path for the money supply or some similar rule, would—so the argument goes—eventually change the pattern of expectations and eliminate the stickiness.

A large and rapidly expanding body of recent research on implicit contracts, principal-agent relationships, and related subjects has begun to flesh out our knowledge of why wages are sticky. Almost universally the implicit contract and related literature con-

cludes that optimal behavior implies a good deal less wage flexibility in the face of changes in the marginal revenue product of labor than would occur in the spot auction markets of the Walrasian model. But this literature deals with the stickiness of *real and relative wages* in response to shocks of various kinds. It seems to have little to say about the macroeconomic stickiness of *nominal wages*. A micro theory of real wage stickiness may help explain the difficulty of adjusting to sudden large changes in aggregate supply conditions like the two oil shocks of the 1970's. But the more familiar problems facing macroeconomic theory and policy have to do with the ability of the economy to adjust to aggregate demand shocks where nominal wage stickiness is a major barrier to successful adjustment. And here the important question is what, if anything, does the new research imply for the behavior of nominal wages? It is to this subject I want to give my attention.

A road map will be helpful. After summarizing the existing literature, the first section concentrates on *relative wage* adjustments, and argues that under optimal arrangements for the determination of wages, relative wages while not rigid will be sticky. They will adjust only gradually to relative changes in the conditions facing individual firms. The paper then argues that contrary to general opinion, relative wage stickiness necessarily produces aggregate *nominal wage* stickiness; if wages move sluggishly in response to the relative conditions facing individual firms, they will move sluggishly in response to aggregate nominal shocks. Several mechanisms that might produce a flexible wage response to nominal shocks, given sticky relative wages, are considered and rejected—various forms of indexing and rational expectations. Finally the “external” nature of the gains from nominal wage flexibility is invoked against the criticism that attributing cyclical unemployment to nominal wage stickiness implies nonrational behavior on

\*The Brookings Institution, 1775 Massachusetts Avenue, N.W., Washington, D.C. 20036 and Professor of Economics, University of Maryland. The research underlying this paper was supported by a grant from the Ford Foundation. I thank Gardner Ackley, Martin Neil Baily, Barry P. Bosworth, Robert J. Flanagan, Stephen M. Goldfeld, Robert Z. Lawrence, Bruce K. MacLaury, Joseph A. Pechman, George L. Perry, Robert M. Solow, and James Tobin for helpful advice and comments.

<sup>†</sup>Presidential address delivered at the ninety-seventh meeting of the American Economic Association, December 29, 1984, Dallas, Texas.

the part of employers and workers. Macroeconomic shocks will gradually overcome the relative wage stickiness and move the aggregate nominal wage level in the desired direction, but the transitional costs are large.

### I. Some Relevant Features of the Implicit Contract Literature

By now the literature on implicit contracts is so large and so diverse that I cannot do it justice in a brief summary. But it is necessary to sketch out a few elements of that research, paying particular attention to several key features that bear on the relationship between wage behavior and aggregate demand shocks.

The flexible-price, market-clearing model, in which economic agents are price takers and prompt quantity adjusters, would be a useful paradigm in a very particular kind of world. In this world labor and product markets would be characterized by a great deal of homogeneity, and so individual transactions would be carried on in very "thick" markets. There would be no reason for preserving a continuity of relationships between workers and firms, customers and suppliers, lenders and borrowers. Workers would be interchangeable; the marginal revenue product of a particular class of workers would be the same regardless of the firm to which the worker was attached. Either labor effort could be easily monitored or it would be completely proportional to labor hours paid for. And, wherever commitments had to be made for the future, expectations about the important variables entering into the decision could be drawn from fixed stochastic distributions, knowledge of which, in turn, could be derived from the recurrent nature of past events. As well as being unbiased, forecasts could significantly improve on coin-flips.

The world that we are trying to model, however, is in fact different in several fundamental ways. Most importantly, in a substantial part of the economy there are large returns to maintaining the continuity of association between workers and firms. Workers acquire nontransferable, firm-specific skills; and some of the cost of the acquisition may be paid by the firm. They

acquire knowledge about the nonwage attributes of a job through experience on that job. Continuity of association also provides firms with hard-to-come-by knowledge about the reliability and productivity of specific workers. Additional transition, search, and moving costs are incurred by workers and firms when the association is broken. Substantial bilateral monopoly rents to continuity are thus generated among firms and their experienced work force.

Risk aversion introduces another reason for continuity of association. For at least some range of possible variations in the economic environment facing a firm, it is likely to be less risk averse than its workers. But the relevant insurance contracts cannot be traded separately from employment contracts, and so continuity of association becomes a joint product with insurance. Finally, in the real world, labor time is not synonymous with performance. Monitoring worker performance or effort is costly, but the payment of higher than the "going wage" will bind workers to the firms with an incentive to avoid shirking. And the particular workers who "survive" the monitoring are those who have found the premium sufficient to avoid monitorable shirking, a piece of valuable information to the firm flowing from continuity of association.

Realizing the benefits from long-term worker-firm relationships requires, of course, some sort of explicit or implicit agreement between employers and workers on the terms of the relationship, especially with respect to wages and employment opportunities. Several problems in modeling these long-term worker-firm agreements have dominated the recent literature. *First*, when the marginal revenue product of labor changes, how are wages and employment to be adjusted while still preserving contractual relationships? *Second*, since firms are much better able than workers to observe the marginal product of labor and since contingent contracts directly tied to the various states of nature cannot practically be designed, how can contract terms allow for some flexibility in meeting changing conditions without giving employers incentives to provide false information about labor's marginal product? This is the problem of

*asymmetric information* and has given rise to the modeling of incentive-compatible contracts. *Third*, since explicit written contracts between unions and employers cover only a small part of the workers involved in long-term relationships, what keeps either workers or firms from violating the implicit agreement or exploiting their half of bilateral monopoly situations when conditions are favorable to do so? Analysis of this issue has given rise to the concept of a firm's labor market "reputation," or brand-name capital, fear of losing which provides an enforcement constraint. This is the problem of *enforcement*.

Three major strands of the literature on implicit contracts can be discerned, each of which emphasizes one of the rationales for the continuity of association (without necessarily denying the existence of other aspects), and each of which deals somewhat differently with the problems cited in the prior paragraph.<sup>1</sup> The earliest version of implicit contracts emphasized the role of *risk aversion*, firms being either risk neutral or less risk averse than workers to fluctuations in their income.<sup>2</sup> Firms thus offer risk-sharing contracts that improve social welfare relative to spot auction markets. Another body of research stresses *transactions costs* and *asset specificity*—that is, the acquisition of firm-specific skills by workers, and specific and valuable knowledge about each other by both firms and workers.<sup>3</sup> This approach also emphasizes the *asymmetry* of knowledge between firms and workers about the marginal revenue product of labor, and its influence on the nature of the contract. Still a different emphasis is given in the *efficiency wage*<sup>4</sup>

literature to an assumed positive correlation between the level (or the career profile) of wages on the one hand and the "effort" or productivity of workers on the other.<sup>5</sup>

While there are substantial differences among these various approaches, they all conclude that, under optimal contracts, real wages will be smoothed in the face of changes in the marginal revenue product of labor relative to what would be predicted by an auction market. And all of them provide a rationale for the existence of *ex post* Pareto inefficiency and involuntary unemployment.

## II. Some Extensions to Implicit Contract Theory

We can distinguish several categories and subcategories of changes in economic circumstances, the wage response to which must be accommodated by social conventions and informal understandings that we call implicit contracts. Let us consider first some of the implications of contract theory as it deals with the response of wages to changes in the *relative* conditions facing individual firms or labor markets. For that purpose I define relative changes to be those which occur on the assumption that the perceived general level of opportunities facing workers—call it  $\bar{W}$ —remains unchanged in real and nominal terms. (The relevant  $\bar{W}$  is, of course, in real terms. But since we are here abstracting from any aggregate disturbances, real and nominal  $\bar{W}$  are the same.) In turn there are two kinds of relative changes in economic conditions that implicit contracts must allow for; changes which are realizations of a probability distribution known at the time the contract is entered, and those stemming from developments to which no basis could be found for assigning probabilities.

<sup>1</sup>I take the threefold categorization from the summary of wage contracting literature by Robert Flanagan (1984). See also Costas Azariadis and Joseph Stiglitz (1983).

<sup>2</sup>See Martin Baily (1974), Azariadis (1975), Donald Gordon (1974), Sanford Grossman and Oliver Hart (1981) and Bengt Holmstrom (1983).

<sup>3</sup>See Benjamin Klein (1984), Klein, Robert Crawford, and Armen Alchian (1978), David Mayer and Richard Thaler (1979), Michael Wachter and Oliver Williamson (1978).

<sup>4</sup>See Carl Shapiro and Stiglitz (1984), Edward Lazear (1981), and Janet Yellen (1984).

<sup>5</sup>In some of the efficiency wage models, the firm attempts to pay more than the going wage, creating a penalty for workers who shirk and are fired. When all firms do this, aggregate unemployment is created and a shirking penalty still exists. No implicit contracts are needed. Both Victor Goldberg (1982) and Lazear, however, postulate a positively shaped earnings profile (new workers earn less than their marginal product) as a way of getting employees to post a performance bond thereby creating shirking penalties. This version does require implicit contracts enforced on the firm by fear of reputation loss.

To deal with implicit labor contracts covering the long tenures that are typical in U.S. industry, it is necessary to make the currently out-of-fashion distinction between risk and uncertainty. Most of the mathematical modeling of implicit contracts has assumed that workers and firms base their agreements on a *known* probability distribution (presumably commonly held) of the relevant variables, most importantly the marginal revenue product of labor or some related variable. The distribution of the mean expected bilateral monopoly rents from continuity is decided at the beginning of the contract on the basis of the known distribution of possible economic environments that will be faced over the life of the contract, and the contract then specifies the behavior of wages (and in some cases employment) as the realization of that distribution occurs. The wage is either rigid or changes sluggishly in the face of these changing realizations. But recent research on the surprisingly long lengths of job tenure in the United States casts doubt on the usefulness and sufficiency of this assumption. Robert Hall (1980) has estimated that, in 1973, half of all work in America was done on jobs whose completed tenures were fifteen years or more. And for men alone the relevant completed job tenure was twenty-five years! Douglas Wolf and Frank Levy (1984) reached very similar conclusions on the basis of a 1979 survey. The contracts, rules-of-behavior, and social conventions that make possible such long associations must be such as to allow wages to adjust appropriately in response to changes in circumstances whose probability of occurrence could not be determined in advance. In other words, the informal agreements which make possible long-term association between workers and employers must take into account Knightian uncertainty about future possible outcomes. With respect to many of possible states of the world, over such a long period of time, there is no basis in past statistical regularities for knowing the distribution. As William Nordhaus (1976) points out, contracts must take account of changes in the economic climate (i.e., when the parameters of the distribution shift) as well as changes in the economic weather (i.e., as realizations of the known distribution).

Most of the possible long-term changes to be faced by firms and workers do not result from the cyclical variance of aggregates, like national income and output, but from changes in relative variables potentially responding to a bewildering permutation of possibilities. Seen in 1973, what were the rationally expected probabilities of the 1974 and 1979 oil price increases, the introduction and growth of personal computers, or the Chrysler brush with bankruptcy? Compared to workers, firms may indeed be relatively risk neutral to temporary changes in income following some known distribution. But no firm is so risk neutral and has such unlimited access to capital as to enter upon or honor contracts specifying rigid wages or some fixed function of wages on employment over very long periods,<sup>6</sup> when no rational basis exists for specifying the distribution of outcomes. Some of the modern wage literature (for example, Hall and David Lilien, 1979; Sanford Grossman and Oliver Hart, 1981) models a contract with a "lump-sum" distribution of the rents—that is, an amount to be paid the workers regardless of unemployment status—and a marginal compensation, paid when the worker is employed and itself an agreed upon increasing function of the level of labor input. Employers then determine employment by maximizing profits in the light of the marginal compensation schedule. In these approaches, the climatic changes to which I refer would be an occasion for changing the basic lump sum distribution of the rents.

A workable distinction can thus be made between those contract provisions which deal with wages in the face of moderate and temporary changes in the marginal product of labor that are perceived as the realization of a known probability distribution and those provisions which deal with climatic and permanent changes, the probability of whose

<sup>6</sup>Indeed, for contracts covering long periods of time, workers are probably less risk averse than firms. Under some range of unfavorable outcomes, rigid wages could mean bankruptcy for the firm. Against relative changes in fortunes the worker loses only his "rent" from continuity plus search costs. And, however well stockholders may be able to diversify against bankruptcies, the managers of the firm find it much more difficult.



occurrence cannot be estimated before the fact. With respect to changes in labor demand that are perceived to be consistent with a previously known distribution, implicit contract theory predicts either rigid wages or—in the models like those of Hall and Lilien and Grossman and Hart—wages which move less than spot auction markets would predict but are positively correlated with the level of employment. But the informal agreements or generally accepted conventions that we call implicit contracts must also provide for changes in circumstances whose probability cannot rationally be estimated at the time workers enter into the contract. These climatic changes in economic circumstances can be of several broad kinds. They might involve a relatively long-term change, because of shifts in demand or costs, in the rents available to be shared by a firm and its workers. Long-term changes in the relative demand and supply of particular occupations or in particular local labor markets will also call for changes in relative wages around a given  $\bar{W}$ .

If the present value of the stream of benefits to workers (wages and employment probabilities) flowing from continuing with the firm begins to decline relative to the alternative combination of initial investment costs and future rents that can be generated at other firms, it is appropriate that contract terms signal the information to workers, so that they can make the relevant comparisons, and decide whether to quit the firm. Similarly, if a substantial and permanent decline occurs in the relative demand for labor by occupation or locality, wages ought to signal the change. Conversely, economic circumstances might increase the rents to be divided making it optimal for the firm to enlarge its share of the relevant labor market pool of experienced workers. In that case, the signal of higher relative wages ought to be transmitted to attract from other firms or occupations workers for whom the lost rents are less than the improved opportunities.

An efficient determination of wages must thus cope with two hard-to-reconcile sets of facts. Because long continuity of association between firms and specific workers confers substantial benefits on both parties, rent-sharing arrangements that offer protection

against exploitation and promote such continuity are profitable. But the period of association is so long that the probability distribution of circumstances requiring changes in wages is largely unknowable. And so the rules and conventions governing wage determination must be flexible enough to allow response. Seen in this light, the terms "implicit contract" may be misleading. It is, I think, useful to think of wages as being set and periodically revised by firms within limits imposed by a set of social conventions, concepts of equity and fairness, and informal understandings. These conventions, concepts, and understandings provide substantial, if imperfect, protection against exploitation but allow flexibility to deal with Knightian uncertainty.

While very lengthy association requires relative wage adjustments in response to "unforeseeable" changes in circumstances, other characteristics of the labor market suggest that those adjustments will tend to be gradual and delayed. It would be hard to account for the long job tenures that we observe unless the combination of positive returns to association and the cost of transition were quite sizeable. The quantitative evidence on the magnitude of explicit turnover costs suggests that they are large. Daniel Mitchell and Larry Kimbell (1982), for example, report a recent estimate, based on a survey of Los Angeles firms, that firms' own costs of turnover (exit costs plus replacement costs) averaged \$3,600, \$2,300, and \$10,400 for production, clerical, and salary-exempt workers, respectively.

Under these circumstances, erroneous signals can have asymmetrical results. Changes in wages undertaken on the mistaken identification of a temporary change in circumstances as a permanent change can cause a substantial loss to workers and firms from the unnecessary scrapping of "investment" and the incurring of other turnover costs. Entire rent streams are wiped out to the extent that workers having transferred to other firms cannot return when the mistake is discovered. Moreover, once experienced workers do change jobs, it is likely to take several tries before they find another long-tenure job. Hall (1982), for example, shows that in 1978 a worker aged 45–49 *who was in*

a new job had only a 20 percent probability of holding that job as long as five years.

While the large magnitude of rents for experienced workers tends to provide some room for errors, workers are presumably arranged along a spectrum with respect to their own evaluation of potential opportunities elsewhere. As a consequence, the losses from the errors will be a continuous positive function of the size of the errors. Errors of the opposite type—failing to introduce a contract change to meet a permanent alteration in the climate—can be reversed at a much smaller cost. Some workers will have remained too long with a particular firm, while a smaller group will have lost income from the higher layoffs associated with this type of error. But these losses are likely to be much smaller than the loss of the rent stream itself, unless the error persists for a long time.

Thus, given Knightian uncertainty about economic conditions over the long length of job tenures, implicit contracts must allow for the possibility of relative wage changes in the face of climatic changes in external conditions. But the interest of both firms and workers dictates that those changes occur only after enough information has been accumulated to warrant a high probability forecast that the change is permanent. To justify a substantial wage adjustment, it is not sufficient that the firm act on unbiased forecasts—they must also acquire some confidence in the accuracy of the forecasts.

One might object to this line of reasoning on grounds that failure to adjust wages quickly enough while the evidence is accumulating that the change in circumstances is permanent, is itself likely to send out wrong signals. If, for example, wages are slow to adjust downward when the demand for labor falls, employment will decline. Why will workers not take this as a signal that the future stream of benefits from staying in the current job have declined relative to earlier anticipation? There are a number of reasons why this objection is not valid. First, a change in wages is known immediately, while it takes some time to begin to realize that an actual change in employment reflects a shift in the long-term distribution of employment prob-

abilities. Second, the common practice of layoffs subject to recall is a way for a firm to signal that the lower employment is expected to be temporary. Third, significant downward changes in relative wages are not lightly made. Given the inability of workers easily to assess the magnitude of the available rents, firms are deterred from exploiting the bilateral monopoly relationship through the damage they might do to their "reputations" and the future increases in employment costs thereby imposed. Since reputation is a fragile asset, and since workers are naturally less likely than firms to interpret current facts as warranting a relative wage cut, firms must wait until a substantial body of evidence points in the required direction. The efficiency wage literature emphasizes the unfavorable productivity consequences of relative wage reductions that turn out to be unwarranted. And, since firms know that it is difficult to reduce wages once they have been raised, they do not want to make a mistake in the upward direction. As a consequence, wage changes that survive these barriers are much more clearly interpreted as a signal that the stream of future rents has changed than are variations in employment.

Efficient implicit wage contracts, in the presence of uncertainty, must also have the characteristic that they minimize "haggling costs." They should not lead to a constantly renewed battle over the division of the rents. Frequent struggles would waste resources, possibly reduce productivity, and erode scarce reputation capital, thereby reducing the probability of long-tenure associations. These considerations argue that significant relative wage adjustments to meet perceived permanent changes in condition be an infrequent occurrence and not lightly changed once made.

In sum, modern contract theory concludes that relative wages will tend to be quite sticky in comparison to predictions from the auction market model, in the face of changes in conditions that are the realizations of known probability distributions—what I have loosely called temporary changes. But the existence of very lengthy job tenures and the large gains from continuity suggest that the social conventions and informal agree-

ments which we call implicit contracts must provide for adjustments in wages that move towards market clearing in response to unforeseeable permanent changes in relative economic conditions. Finally, however, the substantial penalties which can be suffered if firms and workers respond to erroneous signals by prematurely severing association call for informal arrangements and agreements that produce a very slow and cautious adjustment of wages even to what ultimately turn out to be permanent changes in relative conditions.<sup>7</sup>

It is not surprising, given the large social returns to continuity of association and the very great difficulty of distinguishing temporary from permanent changes in economic circumstances, that society should have developed social conventions and informal agreements that minimize the sending of premature signals for a reallocation of resources.

### III. Response to Aggregate Demand Shocks

So far we have considered a world in which only relative or local changes are permitted, a world in which the nominal and real value of the general wage level, or the wage "norm," was fixed. Now impose aggregate demand shocks on such a world, optimal adjustment to which requires a change in the path of average prices and

wages. (For simplicity of exposition we will be considering only aggregate demand shocks, and exclude aggregate shocks to the supply curve, like those arising from OPEC-imposed oil price increases or from crop shortages. We can thus assume that the equilibrium solution will not call for changes in aggregate real wages.) An efficient system of implicit contracts must obviously provide for adjustments in the nominal wages of individual firms when changes occur in the average level of nominal wages,  $\bar{W}$ . Under the social conventions and understandings which govern wage determination, there is a rebuttable presumption that—barring the existence of circumstances which call for changes in real and relative wages—nominal wages in each firm will be adjusted in line with observed changes in prices and wages generally. But this process of wage determination does not generate prompt and flexible adjustment of aggregate nominal wages to nominal shocks. A change in the average level of wages is the product of changes in wages by individual firms. To the extent that individual wage adjustments must wait on changes in the average, aggregate nominal wage flexibility will *not* be a characteristic of the system. And, as discussed below, this is also true, but to a somewhat lesser degree, of the wage response to price changes. Given the substantial costs of sticky nominal wages to society as a whole and to individual firms and their workers, is there not some other adjustment mechanism which would generate a prompt and flexible nominal wage response, preserving relative wages but producing the desired nominal flexibility?

Several lines of inquiry suggest themselves. *First*, why are not nominal wages explicitly indexed to some aggregate nominal indicator, producing the appropriate change in  $\bar{W}$  in response to aggregate demand shocks? *Second*, even with relative wage stickiness, would not rational firms and their workers forecast the ultimate equilibrium change in  $\bar{W}$  and promptly set individual wages accordingly? And, *finally*, if that is not feasible, why do implicit contracts not permit swifter and larger nominal wage adjustments by individual firms under the force of aggregate nominal shocks?

<sup>7</sup>Ian McDonald and Solow (1981) have examined union wage bargaining, in which the bargain determines both employment (for example, via work rules) and wages. They assume symmetric information and show that under a number of different bargaining conventions—including one which preserves for the parties "fair shares" in net revenues—shocks to the firm's demand curves may produce small or zero change in wages, so long as  $\bar{W}$  is constant. Some strands of implicit contract literature also model the contracts as specifying certain aspects of the employment decision. Employers do tend to hoard labor during downturns. The implicit contract can thus be viewed as specifying a wage and, for a known distribution of future conditions, a probability of suffering unemployment. Applying the McDonald-Solow analysis to this arrangement, one could conclude that even unforeseen "climatic" changes in circumstances facing the firm would more likely result in a change in the unemployment probability than in the wage.

Let us start with the issue of indexing wages in individual firms to the general price level. In the United States, wages are not widely protected against changes in price by explicit indexing formulae. In 1983, only 58 percent of union workers were covered by COLAs. On the average, the ones that were, received protection against only 53 percent of the changes in the *CPI*. Jo Anna Gray (1978) and Stanley Fischer (1977) examined the conditions under which indexing would or would not be optimal for individual firms given the assumption that wages once set are not renegotiated for some period. The main conclusion from this research is that, in the face of real (as opposed to nominal) shocks, indexing, by freezing real wages, produces real wage results that are not optimal for the firm. Given the probability that nominal and real shocks are both likely to occur, Gray shows that *partial* indexing will be an optimal choice. Moreover, even nominal shocks are likely, during the transition to a new equilibrium, to have nonneutral effects; the firm's product price and the Consumer Price Index may not move in parallel. As a consequence, indexing would have unwanted real effects even in the face of monetary shocks. While we have seen that some degree of real wage stickiness is optimal, the absence of full indexing even in multiyear union contracts indicates that the opposite extreme—automatic guarantees of a fixed real wage over several years—is not a workable approach.

Within-year indexing is virtually nonexistent in annual union contracts and in the annual wage adjustment cycle followed by the vast bulk of nonunion firms. Under implicit contracts, changes in the path of average wages and prices in the economy as a whole or in relevant submarkets are commonly agreed to constitute a major element in determining the size of those annual wage adjustments. But, in the United States, it is almost universally the practice not to make the relationship to the price level or to average wages an explicit and automatic one, either within the year or over longer periods. And, as noted earlier, in those multiyear union contracts where indexing is found, that indexing is almost always less than complete. Haggling costs are apparently minimized by

making periodic wage adjustments that simultaneously take into account nominal, real, and relative factors, rather than by fixing the nominal relationship in a formula and separately adjusting for changes in real and relative conditions.

In any event, even if wages were fully indexed to prices, they would not produce highly flexible nominal wages in response to aggregate demand shocks. The cumulative costs of wage indexing, in the face of real shocks and the transitional nonneutrality of nominal shocks, would be severely exacerbated if the indexing were instantaneous. As a consequence, the indexing we do observe—except in countries which have developed extremely rapid and sustained inflation—typically involves a substantial lag between observed price changes and wage adjustments. Such indexing only guarantees a gradual response of wages to aggregate demand shocks in proportion as prices themselves are flexible in the face of constant wages. Even if prices were competitively determined, the economy would still have to work its way through a series of price-wage-price reactions in each one of which prices fell, relative to wages, by an amount depending on the slope of the marginal cost curves. And prices do not move with such frequency or flexibility. Arthur Okun (1981) has carefully elaborated the reasons why, in the customer markets which predominate in modern economies, prices are not likely to move quickly and easily up and down a marginal cost curve. Robert Gordon (1981) has elaborated how the highly articulated input-output relationships of modern economies tend to slow the price reaction to aggregate demand shocks. And George Akerlof and Janet Yellen (1984) have recently shown that in less than perfectly competitive markets inertial price-setting behavior in response to a shock may impose only second-order losses on the firms who follow such behavior, even though the macro result may be first-order losses for the economy.<sup>8</sup> Everything else

<sup>8</sup>Technically the Akerlof-Yellen proposition applies to situations in which agents' objective functions are differentiable in their own prices and wages—a condition that is not met in the competitive model, but is met in a wide range of other market models.

being equal, indexing wages to prices would provide some nominal wage flexibility in the face of nominal shocks. But if wages are otherwise sticky, indexing them to prices would still yield a very gradual iterative process of demand inflation or disinflation.

Granted the obstacles to indexing wages to prices in implicit contracts, and the insufficiency of that arrangement—even if feasible—to produce prompt nominal wage adjustments, why are not implicit contracts indexed to some aggregate like nominal *GNP* or the money supply? Such an arrangement might seem to be a way to approximate the role of the Walrasian auctioneer, automatically generating  $\bar{W}$  at a level to clear the aggregate supply-demand balance, while relative wages continued to be set under implicit contracts along lines suggested earlier. In fact, of course, we observe no such arrangement anywhere in the world, and a little thought supplies a number of reasons. For purposes of indexing wage contracts to nominal *GNP*, some agreed-upon process would have to be found for separating “disturbances” in nominal *GNP* from the trend increases consistent with full employment at a stable inflation rate. Anything that altered the parallel growth of average and marginal labor productivity, or the growth of full employment labor inputs, would change the trend and require the contracts to be renegotiated. Robert Gordon (1981, 1983) has identified a number of reasons, why, even if the appropriate split could be made between trend and disturbances, indexing wages to nominal *GNP* would not be feasible in implicit contracts. If prices themselves are not completely flexible relative to wages, declines in nominal *GNP* under an indexed system would produce long periods with unwarranted real wage decreases. And, since the costs and hence the prices of the typical large firm depend on the costs and prices of a long and heterogeneous chain of suppliers, indexing wages on nominal *GNP* would only index part of an individual firm’s costs. Workers would rightly be skeptical that such an indexing system would quickly move prices down proportionally with wages.

The problems of automatic indexing to some other nominal aggregate, like the mon-

ey supply, are even worse since the relationship between any other aggregate variable and the equilibrium full employment wage level is still more complex and unstable than it is in the case of nominal *GNP*. And unless all firms could somehow agree on a common translation formula for indexing purposes, nominal demand shocks would produce a wide dispersion of unwarranted changes in relative wages. (This point is elaborated further in the paragraphs that follow.) More complex explicit indexing formulas can be imagined, but are no more feasible than simple ones. The two parties to a contract would be subjecting themselves to very great uncertainty in agreeing to a given information set and forecasting model as the basis for the indexing. It takes a very large run of data to sort systematic error from noise in economic time-series. And so agents would have huge space for disagreement about the appropriate information set and the relevant model for translating information into forecasts. Alternative choices could lead to biased outcomes, whose bias could not be determined for a very long time. (The appropriate order in which to enter variables in a vector autoregression model is hardly the subject for fruitful labor negotiations.) In short, feasible state-contingent contracts cannot be designed to replace the Walrasian auctioneer as a means of coordinating the system’s response to nominal shocks.<sup>9</sup>

In the absence of explicit state-contingent contracts, can nominal wage flexibility be rescued by a rational expectations model of wage determination? Why do not individual firms and their workers rationally forecast the change in the equilibrium path of average wages ( $\bar{W}$ ) expected to result from a nominal shock, and promptly change wages accordingly, recognizing that their actions involve no decision to change relative wages?

<sup>9</sup>Under extreme circumstances the parties may introduce such contacts. In the German hyper-inflation of 1923, the pace of inflation became so swift that weekly wages were finally indexed to the prices *expected* to prevail in the subsequent days when the wages would be spent. An *ex post* adjustment was then made to correct for errors in the forecast. A similar arrangement was adopted during the Polish hyper-inflation of the same period.

Stochastic errors in forecasting could generate temporary wage "errors" and departures of employment from its natural rate. But nominal wages would fundamentally be flexible.

A rational expectations approach to the determination of the aggregate nominal wage and price level, however, cannot simply be carried over into a world of sticky *relative* wages. There are several reasons why this is so. In the first place, in the new classical model it is *not* the expectational element that generates aggregate nominal wage flexibility. Rather, a prompt response of aggregate nominal wages to nominal shocks is guaranteed by the perfect *ex ante* flexibility of relative wages in auction markets. In these models a downward nominal demand shock generates a prompt and "neutral" change in the path of nominal wages precisely because of individual workers' presumed willingness to underbid wages, shading their bids and offers in the face of excess supplies, to the point where markets are cleared. In the auction-market model, workers or groups of workers are willing to accept lower wages in the face of excess supply even when errors in expectations lead them to misperceive the entire wage cut as a relative one. In the event of such misperceptions, some labor supply is withdrawn but nominal wages still fall. And as soon as the misperception is corrected, wages quickly fall by the remaining amount necessary to eliminate excess labor supply and clear labor markets. In the world of implicit contracts I have described earlier, however, the absence of substantial relative wage flexibility eliminates this basis for prompt nominal wage flexibility.

In the absence of auction markets with their relative wage flexibility, the achievement of aggregate nominal wage flexibility becomes a prisoner's dilemma problem. But, since wages are not completely rigid under implicit contracts, it is not possible that individual firms and their workers forecast the ultimate equilibrium response of average wages to nominal shocks, promptly adjust their own wages to it and thereby solve the prisoner's dilemma in favor of nominal wage flexibility? I think not. Two key features of

implicit contracts interact with one central feature of rational forecasts to reduce sharply the feasibility of commonly shared expectations about the equilibrium  $\bar{W}$  as a surrogate for Arrow-Debreu contingent claims contracts. *First*, firms are wage setters, not wage takers; *second*, frequent haggling over wage changes is very costly to the development of mutually beneficial long-term relationships between workers and firms; and *third*, forecasts of equilibrium wage and price levels are subject to substantial stochastic error and forecast outcomes are likely to be widely distributed over the population of wage-setting firms.

In a world of auction markets, the fact that forecasts of individual agents are widely distributed around the "true" mean is for most purposes irrelevant. In his seminal article (1961), John Muth pointed out that cross-sectional differences in expectations posed no problem for the theory because their aggregate effect would be negligible so long as deviations from the rational forecast by individual firms were not strongly correlated with each other. In auction markets, specific prices or wages are not determined by individual forecasts. But, in a world of implicit contracts with firms as wage setters, they would be. The dispersion of individual forecasts concerning the equilibrium nominal wage  $\bar{W}$ , and therefore the dispersion of individual wage decisions, would often be a wide one. Robert Lucas' words about the role of rational expectations in shaping the actions of economic agents are relevant in this regard:

Neither will [rational expectations] be applicable in situations in which one cannot guess which, if any, observable frequencies are relevant: situations which Knight called 'uncertainty'. It will most likely be useful in situations in which the probabilities of interest concern a fairly well defined recurrent event, situations of 'risk' in Knight's terminology. [1977, p. 15]

Changes in the path of nominal wages, however, have not simply, or even primarily, been driven by recurrent patterns of endoge-

nous events, at least in recent years. The Vietnam War, two massive oil shocks, the introduction of floating exchange rates, and the institution of a new monetary regime in the United States in 1979 have dominated events. Even if the vast majority of firms held a broadly similar view of the way the economic world works, the very great macro-economic uncertainty and the stochastic variance of prior forecasts around the actual outcomes would guarantee a wide dispersion of individual forecasts whenever large shocks occurred.

A wide dispersion of the forecasts of individual economic agents and the experience of errors in prior forecasts would have a number of consequences. In the first place, given the large room for disagreements about the forecast, we have to rule out on moral hazard grounds implicit contracts under which workers, prior to actually observing a deterioration in employment conditions, would accept employer forecasts of a declining equilibrium nominal wage as the basis for a downward wage adjustment. But even waiving this difficulty, the large dispersion of individual forecasts would result in widespread unintended relative wage changes, even if all firms and their workers accepted a wage adjustment based on the firm's equilibrium forecasts. Because of both risk aversion and the consequences to firms and workers that follow from erroneous signals, these changes could impose significant losses on the parties. Yet, a high frequency of wage changes is also costly, so that errors would tend to persist for some time. Indeed, given the incompleteness and imperfections of current information, and the murkiness of the variable being forecast—the equilibrium nominal wage level—there would be substantial room for disagreement among the parties as to whether or not a prior forecast had or had not been in error. Finally, to the extent that “bad” experience with forecasts caused some firms and workers to reduce the forward-looking element in wage setting, the unreliability of forecasts for those who used them would become greater. It would become increasingly less rational to base individual wage decisions on the assumption that

others were forecasting the equilibrium adjustment and promptly adjusting wages accordingly. Thus, without the “policing” mechanism of *relative* wage flexibility, a system that relied on rational expectations forecasts of the equilibrium wage outcome would be unstable.

In sum, under implicit contracts, widespread tying of individual wage decisions to expectations about equilibrium aggregate wage outcomes is not a feasible way of anticipating the optimal adjustment to nominal shocks. The policing mechanism of *ex ante* relative wage flexibility is absent, so that nominal wage flexibility becomes a prisoner's dilemma problem. And prompt rational expectations “indexing” to the equilibrium outcome is no solution to that problem, since it would increase haggling costs, produce unwanted relative wage changes, and become increasingly an irrational action on the part of individual firms.

In the absence of widespread indexing to some nominal aggregate, or to the rational expectation of the equilibrium wage norm  $\bar{W}$ , wages must find their way to a lower level as the product of specific decisions among individual firms and their workers, in a process constrained by the same social conventions and informal agreements that dictate the change in relative wages under implicit contracts. In a world of price and wage setters, firms and workers observe demand shocks principally in the form of changes in their own physical quantities—sales first and then output and employment—and in the context, initially, of an unchanged perceived level of  $\bar{W}$ . The dynamics of the process by which firms adjust from one level of the work force to another can, as described by Okun and by George Perry (1980), generate modest wage changes in response to the aggregate shocks. Beyond this, the change in external pressures for wage adjustment must be large enough and last long enough to satisfy the constraints imposed by long-term implicit contracts on “permanent” wage changes. Finally, to the extent that these changes become substantial and widespread, enough to yield a long-term change in the perceived level of  $\bar{W}$ , the whole nominal

wage structure around which individual adjustments occur will be changed.

There are several points to note about this process. The major gains, in terms of higher employment, that came from lowering wages in the face of downward aggregate demand shocks do not accrue to particular workers in particular firms as the result of their own actions. Rather, the gains accrue through the effect of generalized lower wages in reducing prices, raising real money balances and thereby increasing aggregate demand. Once the assumption of flexible auction markets and a competitive bidding down of wages by individual workers is abandoned, it is no longer valid to level against sticky-wage theories the charge that they imply an irrational failure on the part of economic agents to pursue unexploited gains from trade. Without the Walrasian auctioneer, the individual firms in an economy are, as noted earlier, in a prisoner's dilemma. The large comparative statics gain from an aggregate nominal wage cut does not translate into such a gain seen from the point of view of individual firms.<sup>10</sup> The potential gain they see is the one that would accrue from making a relative change. They will indeed make such changes, but slowly and cautiously, acting under the constraints on relative changes spelled out earlier.<sup>11</sup> Only as the perceived long-term level of  $\bar{W}$  falls will the situation be different.

<sup>10</sup>In his 1977 article, Robert Barro decouples the nominal wage and employment responses to perceived nominal shocks. He argues that even if nominal wages are sticky, optimal contracts would call for the maintenance of employment. Firms would vary prices to achieve this result despite the stickiness of nominal wages. Otherwise, says Barro, the parties would be ignoring the unexploited gains from trade, an irrational act. This paper is not principally concerned with price behavior. But it is clear that the same externality argument, set forth above with respect to wages, holds for prices. Without auction markets or some surrogate for the Walrasian auctioneer, the external nature of the gains from price changes would tend to negate the force of the Barro argument.

<sup>11</sup>Since the macroeconomic effects of their own wage and price decisions do not enter into economic agents' objective functions, the Akerlof-Yellen conclusions about the small size of losses from "near-rational" behavior would apply to the nonauction markets I am here describing.

Since what is important for nominal wage adjustment is the *perceived* level of  $\bar{W}$ , expectations about its future value will, of course, be relevant. Within the constraints imposed by implicit contracts, wages in individual firms have to be adjusted to deal with changing conditions. Since wage changes are difficult and impose strains on long-term relationships, the wage once set has to last for a while, typically at least a year (and under union contracts often longer). The expectations that workers and employers have about the future course of average wages and prices will therefore exert an important influence over the current wage decision. I do not want here to join the controversy over the extent to which wage setting is backward or forward looking. But what is central to my message is that the relevant forecast does not assume prompt adjustment to a new equilibrium wage but rather the more hesitant and gradual process described above.

The appropriate framework for analyzing the aggregate behavior of wages, therefore, is *not* one in which certain macroeconomic imperfections—information gaps or misperceptions about the general wage and price level—prevent the market-clearing adjustment of wages, which themselves are perfectly flexible in the face of relative disturbances. The essence of the behavior of wages in response to aggregate demand shocks is just the opposite. Macroeconomic shocks break through the short-run stickiness of wages and prices in the face of relative disturbances to produce the aggregate adjustments, albeit slow and gradual ones, that we do in fact observe.

There is a nice paradox in all of this. A prompt adjustment of nominal wages to aggregate demand shocks, leaving relative wages unchanged, can be produced only by a system of highly flexible relative wages. Conversely, the existence of sticky relative wages yields long transitional periods in which firms are adjusting individually to nominal shocks, and produces, as a side effect, changes in relative wage and prices.

#### IV. Final Reflections

The large costs which accompany disinflation arise from the fact that society has to



send out the same kind of initial signals—changes in the volume of sales—when it wants a reallocation of resources as it does when it wants a change in the general level of wages and prices. In the first situation, given the substantial efficiencies which flow from long-term associations of suppliers with customers and firms with workers, very cautious and sluggish changes in wages and prices are the optimal response to signals. A large part of the adjustment is optimally taken up as temporary variations in slack—adjustments in hours, layoffs and rehires, inventory building and depletion, and rationing of various kinds. Indeed, a large part of economic life is dominated by the social conventions, institutions, and patterns of behavior that have evolved to avoid the chaos and inefficiencies that would result from continuous market clearing.

The signals which firms initially receive when aggregate demand shocks occur are the same as those for a resource transfer, but an exactly opposite response is wanted—large changes in wages and prices and small changes in quantities or slack. Since the bulk of the disturbances to which individual firms and workers must adjust are relative or local in nature, and since over long periods of time micro efficiency tends to outweigh aggregate resource utilization as a source of economic welfare, wage- and price-setting institutions have developed with a bias toward the sluggish response called for by considerations of micro efficiency. The cyclical behavior of the aggregate wage and price levels unfolds from the gradual overcoming of that bias.

In the long run, those features of economic relationships which make short-run price and wage stickiness optimal and which rationally prevent continuous market clearing disappear. Specific assets are converted to capital. Attrition and learning change the mix of skills. Random changes get smaller compared to systematic changes. The private returns from specific customer-supplier and worker-firm attachment shrink relative to the returns from making appropriate adjustments to changes in tastes, technology, and other external developments. The rationally based barriers to market clearing crumble.

We economists are indeed correct to insist on the long-run efficacy of markets and the utility of the market clearing paradigm as a way of explaining long-term market allocations. But we need not abandon the premise of the rational maximizing calculus in order to explain the structural stickiness of wages and prices and the failure of markets to clear in the short run. Both phenomena—long-run market clearing and short-run stickiness—ultimately derive from the same rational aspects of human behavior.

Some of the consequences of this recognition are nevertheless very troubling for economic theory and theoretically informed empirical research. In the new classical economics, there is no need for empirical research to determine how wages and prices respond to demand shocks, *given* expectations about the general price level. Pure theory—the auction-market model—dictates how prices and wages behave. Empirical research is needed principally to tell us something about the formation of expectations on the general price level.

While contract theory and related research has been developing rationally based foundations for structural wage and price stickiness, the work to date is essentially in the form of existence theorems. That is, it tells us *why* sticky wages are consistent with the rational calculus. But it does not give us a theoretical basis for specifying the two basic components of macro wage adjustment: What “laws” do firms follow, under implicit contracts, in adjusting their wages, assuming the stability of  $\bar{W}$ , the wage norm? And what does it take to produce a perceived “permanent” change in that norm?

A full and complete microeconomic foundation to wage adjustment with the power of the auction-market model may never be forthcoming. If that is so, we may have to look to regularities derived from macroeconomic empirical research to infer microeconomic behavior. But this raises another set of problems. While forward-looking expectations play much less of a role in the macro wage adjustment process I have outlined, they are not completely absent in forming perceptions about the wage norm. Hence the force of the Lucas critique, though weakened,

does not disappear. The new classical economics simply assumes that structural behavior is market clearing, and hence claims to be able to identify the expectational effects of a particular policy regime. In the absence of such an a priori assumption, however, how does one go about separately identifying the expectational from the structural elements in wage formation? I do not have the answer. Conceivably, economics, like physics, is subject to a fundamental indeterminacy theorem.

### REFERENCES

- Akerlof, George A. and Yellen, Janet L., "A Near-Rational Model of the Business Cycle with Wage and Price Inertia," unpublished, May 1984.
- Azariadis, Costas, "Implicit Contracts and Underemployment Equilibria," *Journal of Political Economy*, December 1975, 83, 1183-202.
- \_\_\_\_\_ and Stiglitz, Joseph, "Implicit Contracts and Fixed Price Equilibria," *Quarterly Journal of Economics*, Supplement 1983, 98, 1-22.
- Baily, Martin N., "Wages and Employment Under Uncertain Demand," *Review of Economic Studies*, January 1974, 41, 37-50.
- \_\_\_\_\_, "Comment," *Brookings Papers on Economic Activity*, 1:1980, 125-32.
- Barro, Robert, "Long-Term Contracting, Sticky Prices, and Monetary Policy," *Journal of Monetary Economics*, July 1977, 3, 305-16.
- Bresciani-Turroni, Constantino, *The Economics of Inflation*, London: George Allen and Unwin, 1937.
- Carlson, John A., "A Study of Price Forecasts," *Annals of Economic and Social Measurement*, Winter 1977, 6, 27-56.
- Fischer, Stanley, "Wage Indexation and Macroeconomic Stability," in Karl Brunner and Allan Meltzer, eds., *Stabilization of the Domestic and International Economy*, Vol. 5, Carnegie-Rochester Conferences on Public Policy, *Journal of Monetary Economics*, Suppl. 1977, 107-48.
- Flanagan, Robert J., "Implicit Contracts, Explicit Contracts, and Wages," *American Economic Review Proceedings*, May 1984, 74, 345-49.
- Goldberg, Victor P., "A Relational Exchange Perspective on the Employment Relationship," Working Paper Series No. 208, Department of Economics, University of California-Davis, October 1982.
- Gordon, Donald F., "A Neoclassical Theory of Keynesian Unemployment," *Economic Inquiry*, December 1974, 12, 431-59.
- Gordon, Robert J., "Output Fluctuation and Gradual Price Adjustment," *Journal of Economic Literature*, June 1981, 19, 493-530.
- \_\_\_\_\_, "A Century of Evidence on Wage and Price Stickiness in the United States, the United Kingdom and Japan," in James Tobin, ed., *Macroeconomics, Prices, and Quantities*, Washington: The Brookings Institution, 1983, 85-133.
- Gramlich, Edward M., "Models of Inflation, Expectations Formation: A Comparison of Household and Economist Forecasts," *Journal of Money, Credit and Banking*, May 1983, 15, 155-73.
- Gray, Jo Anna, "On Indexation and Contract Length," *Journal of Political Economy*, February 1978, 86, 1-18.
- Grossman, Sanford J. and Hart, Oliver D., "Implicit Contracts, Moral Hazard, and Unemployment," *American Economic Review Proceedings*, May 1981, 71, 301-07.
- \_\_\_\_\_ and \_\_\_\_\_, "Implicit Contracts Under Asymmetric Information," *Quarterly Journal of Economics*, Supplement 1983, 98, 123-56.
- Hall, Robert E., "Employment Fluctuations and Wage Rigidity," *Brookings Papers on Economic Activity*, 1:1980, 91-124.
- \_\_\_\_\_, "The Importance of Lifetime Jobs in the U.S. Economy," *American Economic Review*, September 1982, 72, 716-24.
- \_\_\_\_\_ and Lilien, David M., "Efficient Wage Bargains under Uncertain Supply and Demand," *American Economic Review*, December 1979, 69, 868-79.
- Holmstrom, Bengt, "Equilibrium Long Term Labor Contracts," *Quarterly Journal of Economics*, Supplement 1983, 98, 23-54.
- Klein, Benjamin, "Contract Costs and Administered prices: An Economic Theory of

- Rigid Wages," *American Economic Review Proceedings*, May 1984, 74, 332-38.
- \_\_\_\_\_, Crawford, Robert G. and Alchian, Armen, "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process," *Journal of Law and Economics*, October 1978, 21, 297-326.
- Lazear, Edward P., "Agency, Earnings Profiles, Productivity, and Hours Restriction," *American Economic Review*, September 1981, 71, 606-21.
- Lucas, Robert E., Jr., "Understanding Business Cycles," in Karl Brunner and Allan H. Meltzer, eds., *Stabilization of the Domestic and International Economy*, Vol. 5, Carnegie-Rochester Conferences on Public Policy, *Journal of Monetary Economics*, Suppl. 1977, 7-30.
- Mayer, David and Thaler, Richard, "Sticky Wages and Implicit Contracts: A Transactional Approach," *Economic Inquiry*, October 1979, 17, 559-74.
- McDonald, Ian M. and Solow, Robert M., "Wage Bargaining and Employment," *American Economic Review*, December 1981, 71, 896-908.
- Mitchell, Daniel J. B. and Kimbell, Larry J., "Labor Market Contracts and Inflation," in Martin Neil Baily, ed., *Workers, Jobs and Inflation*, Washington: The Brookings Institution, 1982, 199-238.
- Muth, John, "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, 29, 315-35.
- Nordhaus, William D., "Comment," *Brookings Papers on Economic Activity*, 3:1976, 623-27.
- Okun, Arthur, *Prices and Quantities: A Macroeconomic Analysis*, Washington: The Brookings Institution, 1981.
- Perry, George L., "Inflation in Theory and Practice," *Brookings Papers on Economic Activity*, 1:1980, 207-41.
- Shapiro, Carl and Stiglitz, Joseph E., "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review*, June 1984, 74, 433-44.
- Solow, Robert M., "Comment," in James Tobin, ed., *Macroeconomics, Prices, and Quantities*, Washington: The Brookings Institution, 1983, 279-84.
- Wachter, Michael L. and Williamson, Oliver E., "Obligational Markets and the Mechanics of Inflation," *Bell Journal of Economics*, Autumn 1978, 9, 549-71.
- Wolf, Douglas A. and Levy, Frank, "Pension Coverage, Pension Vesting, and the Distribution of Job Tenures," in H. Aaron and G. Burtless, eds., *Retirement and Economic Behavior*, Washington: The Brookings Institution, 1984, 23-63.
- Yellen, Janet L., "Efficiency Wage Models of Unemployment," *American Economic Review Proceedings*, May 1984, 74, 200-05.
- League of Nations, *The Course and Control of Inflation: A Review of Monetary Experience in Europe After World War I*, Washington, 1946.

# The Production Process in a Competitive Economy: Walrasian, Neo-Hobbesian, and Marxian Models

By SAMUEL BOWLES\*

Recent years have witnessed a growing interest in the internal organization of the firm. Many, taking the work of Ronald Coase (1937) as their starting point, have developed insights based on the concept of transactions costs. Others, building on the work of J. R. Commons (1918, 1935), have developed an historical and institutional analysis of the structure of collective bargaining and internal labor markets. Others, starting from Marx's distinction between work ("labor") and labor time ("labor power") have developed an analysis of class conflict within the firm.

A careful reading of this diverse body of literature suggests that there are many common points of reference. All, for example, have stressed the social and nonmarket aspects of the production process.<sup>1</sup> But there are important differences as well.

In this essay I develop an underlying microeconomic logic of the Marxian model, and contrast it with two alternative views. The first is the simple Walrasian model in which the production process is represented as a set of input-output relations selected from an array of feasible technologies by a

process of cost minimization with respect to market-determined prices. The Walrasian model presents no analysis of the internal social organization of the firm.

The second group of models stems from Coase's seminar work, and is exemplified by the important recent contributions of Armen Alchian and Harold Demsetz (1972), Oliver Williamson (1980), Guillermo Calvo (1979), Edward Lazear (1981), and others. Like the Marxian approach, and unlike the Walrasian, these models present a well-developed model of the firm as a social organization. I refer to these models as neo-Hobbesian because according to them the key to understanding the internal structure of the firm is the concept of malfeasance. Also known as shirking or free riding, malfeasance gives rise to the archetypal Hobbesian problem of reconciling self-interested behavior on the part of individuals with collective or group interests. Moreover, the neo-Hobbesian explanation of the functional nature of the hierarchical organization of the modern workplace bears a close resemblance to the original Hobbesian rationale for the state as a socially necessary form of coercion.<sup>2</sup>

By contrast, the basic commitment of the Marxian models is to the fundamental importance of class as an economic concept. While the Marxian model does not deny the importance of the Hobbesian conflict between individual and collective rationality as an underlying social problem central to an understanding of the production process in any social system, it focuses on those problems which may be traced to the structure of

\*Department of Economics, University of Massachusetts, Amherst, MA 01003. I have benefited greatly from the comments and criticisms of my colleagues at the University of Massachusetts and the University of Siena, particularly Robert Costrell, Kenneth Flamm, Herbert Gintis, Richard Goodwin, Donald Katzner, Michael Kruger, Fabio Petri, Ugo Pagano, and Leonard Rapping. I also thank David Gordon, Robert Boyer, Frank Hahn, James Malcolmson, Robert Solow, Juliet Schor, Robert Gordon, Duncan Foley, and two anonymous referees for criticism and comments, and the John Simon Guggenheim Foundation and the German Marshall Fund of the United States for financial support.

<sup>1</sup>The list of approaches is quite partial as it excludes, for example, the interesting and related work on social norms and economic processes. See George Akerlof (1980) and Robert Solow (1980).

<sup>2</sup>Thus, for example, it is argued that a team of workers would rationally hire a supervisor to monitor their work activities, an economic analogue to the Hobbesian position which asserts that uncoerced citizens in a state of nature would in their own interests commit themselves to obey the dictates of a state.

ownership and control of the means of production.<sup>3</sup>

What is at issue between Marxian and non-Marxian economists is not the general relevance of class concepts to the analysis of social groupings, institutions, or political action, but the status of class as an *economic* concept. Even within the realm of economics, terminological differences aside, there is general agreement on the relevance to a wide range of issues of what Marxian economists would term the class structure. Few economists of any persuasion would question the importance of the distribution of the ownership of assets as a determinant of the distribution of income, patterns of consumption, or levels of saving.

The Marxian model is distinct, however, in that it asserts that consideration of the ownership of the means of production, and the command over the production process which this ownership permits, is essential to a coherent analysis of the production process itself, and to the analysis of market equilibration and competition. It is thus not only in its macroeconomic theory and its theory of collective action that the Marxian model makes substantive use of the idea of class, but in its microeconomics as well.<sup>4</sup>

The distinctiveness of the Marxian microeconomics with respect to the neo-Hobbesian and Walrasian approaches, as we shall see, has little to do with the labor theory of value, however. Its primary focus is on the interactions between the voluntary relations of the marketplace and the command relationships of the workplace. Thus Marxian economists take strenuous exception to Paul Samuelson's assertion that "in the competitive model it makes no difference whether capital hires labor or the other way around" (1957, p. 894).

The structure of the Marxian model may be illustrated by reference to three propositions central to its analysis of capitalist production.

First, capitalists (owners of firms or their representatives) will generally select methods of production which forego improvements in productive efficiency in favor of maintaining their power over workers. For this reason, the technologies in use in a capitalist economy, as well as the direction of technical change, cannot be said to be an efficient solution to the problem of scarcity, but rather, at least in part, an expression of class interest. This proposition is fundamental to the Marxian assertion that the productive potential of a society (the "forces of production") is inhibited (or "fettered") by the specifically capitalist institutional structure of the economy (the "social relations of production").

Second, it will generally be in the interest of capitalists to structure pay scales and the organization of the production process to foster divisions among workers, even to the extent of treating differently workers who are identical from the standpoint of their productive capacities. This proposition is central to the Marxian divide and rule interpretation of internal labor markets, segmented labor markets, and discrimination.

Third, involuntary unemployment is a permanent feature of capitalism central to the perpetuation of its institutional structure and growth process. In a capitalist economy, product and labor markets will not function so as to eradicate Marx's familiar "reserve army of the unemployed." Moreover, even public policy towards this objective will be unable to maintain full employment.

To economists trained in the Walrasian or more generally neoclassical tradition, these assertions are often thought to be either nonsensical, or based on a radically different model of production and competition. Specifically, it is often thought that these propositions require one or more of the following assumptions: that capitalists collude in pursuit of their collective interests, that capitalists do not maximize profits, that product and factor markets are not competitive, or that the economy is characterized by im-

<sup>3</sup>I will specify what I take to be the principal differences between the neo-Hobbesian and the Marxian models in the penultimate section. The relationship between the Marxian model and what Marx wrote is suggested in various notes.

<sup>4</sup>I would thus take strong exception to Oskar Lange's (1935) view that the specificity and strength of Marxian economics resides in its institutional and sociological content and not in its microeconomic theory per se.

portant institutional rigidities such as sticky wages. Under these assumptions, it is not difficult to demonstrate the above propositions and thus affirm the importance of the Marxian concept of class.

But, while sufficient, these assumptions are not necessary to the demonstration of the above basic propositions of Marxian economics. (Nor, one might add in passing, are they particularly central to Marx's own theoretical writings, which generally presumed a highly competitive economy based on profit maximization.) The basic difference between the Marxian and Walrasian models is thus not in the structure of markets or in concepts of collective vs. atomistic action, or in institutional rigidities, but in the analysis of the process of production itself, or in what Marxists term the labor process.<sup>5</sup>

In this essay I develop a simple model of the production process in a competitive capitalist economy. To the familiar two-equation Walrasian model of production (production function and cost function), I add a third equation representing class conflict within the production process. I then derive the above three propositions from the expanded model. I close with some observations on the closely related but quite distinct neo-Hobbesian model of the production process.

My intent is not so much to advance the discussion of technical change, discrimination, or involuntary unemployment per se, as to provide a single coherent microeconomic framework capable of integrating important modern Marxian contributions in these fields. To cite only a few: those of Stephen Marglin (1974), William Lazonick (1982), and Harry Braverman (1974), on technology; of Richard Edwards, David Gordon, and Michael Reich (1982), Herbert Gintis (1976), and John Roemer (1979), on divide and rule strategies; and of Michel

Kalecki (1943), Andrew Glyn and Robert Sutcliffe (1972), Raford Boddy and James Crotty (1975), and Richard Goodwin (1967), on unemployment.

### I. The Extraction of Labor from Labor Power

The Marxian model comprises an analysis of three quite distinct aspects of the production process, broadly construed: market exchanges (modeled as voluntary contractual or contract-like interactions), physical input-output relations (which in principal might be represented by an engineering production function), and social relationships among workers and between workers and their employer (which are modeled in an entirely different manner).

Central to the Marxian approach is the distinction between those social relationships that take the form of market exchanges between firms and other ownership units, on the one hand, and relationships of command that take place within firms. The market arena in which contractual exchanges take place, Marx termed "a very Eden of the innate rights of man." By contrast the internal structure of the firm—which Marx termed the "hidden abode of production"—is represented (as Coase was later to do) as a mini-command economy.<sup>6</sup>

The distinction between the two types of social relationships would be of little theoretical importance, of course, if the command relations of the firm were simply effects entirely derived from the technological struc-

<sup>6</sup> The distinction is perhaps the most fundamental in Marxian economics. Marx wrote:

If we consider the exchange between capital and labor, then, we find that it splits into two processes which are not only formally but also qualitatively different...: (1) the worker sells his commodity... (labor power)... which has... as a commodity... a price.... (2) The capitalist obtains labor itself... he obtains the productive force which maintains and multiplies capital.... The separation of these two processes is so obvious that they can take place at different times and need by no means coincide. The first can be and usually, to a certain extent, is completed before the second even begins.... *In the exchange between capital and labor the first act is an exchange and falls entirely within ordinary circulation; the second is a process qualitatively different from exchange, and only by misuse could it have been called any kind of exchange at all.* [1973, pp. 274–75]

<sup>5</sup> Partly as a result of the differing treatment of the labor process and partly for other reasons, the Marxian and Walrasian views of the competitive process differ somewhat. Both stress the importance of unlimited entry and a multiplicity of buyers and sellers. Marxists, however, generally assume price-making rather than price-taking behavior by firms.

ture of production and the market relationships into which the firm enters. Indeed, this is precisely the logic of Samuelson's remark quoted above.

But, according to the Marxian model, the structure and effects of the social relations within the firm—of command, cooperation, competition, and the like—while influenced by technology and market relations, are not entirely reducible to them, but rather depend on the class structure of the productive process, and hence require a distinct form of modeling. By contrast, Walrasian theory denies the need for a distinct modeling of the social relationships within the firm, while the neo-Hobbesian approach insists that a distinct modeling of the firm as a command economy is necessary, but has nothing to do with the class structure, for hierarchical relationships between managers and workers reflect nothing more than an efficient solution to the universal problem of malfeasance.

The importance of the social structure of the firm, the necessity of a distinct modeling of these social interactions, and the centrality of the class structure to their analysis may be traced within the Marxian model to three characteristics of the production process. First, labor is embodied in people, and hence labor services are inseparable from the person supplying the service. Second, whether for reasons of technology or of economies of supervision, production is generally less costly when it is done by a considerable number of workers together in one location. And third, the production process is always a process of joint production, as the workers' attitudes, capacities, and beliefs are transformed in the production process as surely as the raw materials and other goods in process are transformed into final outputs. I will refer to these three characteristics respectively as the human embodiment of labor, the social nature of production, and the endogeneity (or joint production) of workers.

Two types of social interaction within the firm are central to understanding the production process: relations among workers (of competition, solidarity, or whatever) and relations between workers and their employer. I focus on the second at the outset, representing the capital-labor relationship as a

simple bilateral relationship between two individuals. Relations among workers will be introduced later.

The relationship between workers and their capitalist employer is formally structured by the ownership and control of the means of production. It is thus (by definition) a class relationship. In what follows, two characteristics of this relationship will be central. Both may be considered axioms with respect to the proposition to be demonstrated below. First, quite apart from the level of wages, employers and workers have a conflict of interest in the production process in the specific sense that the employer's interests (as measured by profits) are enhanced by being able to compel the worker to act in a manner that he or she otherwise would not choose. This conception of a conflict of interest does not imply that the employer and the worker have no common interests, or that, if left to their own devices, labor would choose not to produce anything at all. It simply states that within a given legal and economic context, the employer can do better than to simply hire workers and let them work as they please. The level of profits therefore depends—at least to some extent—on the power of capital over labor.

While this conflict of interest may extend to such issues as the safety or comfort of the workplace and the amount, type, and location of new investment, I focus in what follows on the conflict over the amount of work done per hour, or what may be termed the intensity of labor. This is often termed the conflict over extraction of labor from labor power. It might better be called the extraction of work from the worker.

The second axiomatic characteristic of the capital-labor relationship is that the strategies that capital may adopt in order to enhance or exercise its power over labor are costly. The basis of the power of capital over labor is the ability of the owner to impose costs on workers who refuse to (or otherwise fail to) carry out the wishes of the employer. In liberal capitalist societies, the only means by which this cost may be imposed is via the employer's control over the terms of employment (wage and other conditions) and the possibility of job termination. For reasons of

simplicity, I focus initially on the threat of job loss.

The expected cost to the worker of resisting (or otherwise not carrying out) the command (explicit or implicit) of the employer will depend on the likelihood that the worker's resistance will be detected, and on the cost to the worker of losing his or her job. (Assume for the present that any worker who is observed performing below the employer's expectation will be fired; I will later modify this assumption.) Because the cost (to the worker) of job loss will depend on the wage, enhancing the threat of job loss (by raising the wage) will be costly to the employer. Similarly, the employer cannot costlessly know what each worker is doing at any given moment even if the employer knows all of the workers' production capacities and personality characteristics. However, the employer can increase the probability of detecting below-standard work intensity through employing surveillance personnel and equipment, and by using production methods that produce (as a joint product) information on individual worker performance. Both methods of enhancing the worker's expected cost of working below expectation are thus costly to the employer.

These two characteristics of the production process—the conflict of interest between capital and labor, and the costliness of employer strategies—form the basis of the propositions that follow. The underlying reasoning may be made more precise with the aid of a simple model.

Let us assume that labor is homogeneous, that the employed and unemployed are otherwise indistinguishable, that there are no employer costs of selection or on-the-job training, that workers are risk neutral, and that all markets are competitive in the sense of a multiplicity of noncolluding buyers and sellers.<sup>7</sup>

<sup>7</sup>Unlike search models or Arthur Okun's (1981) toll model, I assume that workers have complete information about job and wage conditions throughout the economy, that employees know all (actual and potential) employee characteristics, and that what Okun called "the attachment between employer and employees (mutual),...the key component of the toll model that was absent in the simple search model" (p. 75) is absent

Let the output of a firm be a function of the level of inputs.

$$(1) \quad Q = f(X, L),$$

where  $Q$  is the number of units of output over some period of time,  $X$  is the vector of material inputs and services, and  $L$  is the input of labor over this same time period. All inputs and output are measured in physical terms. Labor is thus counted in effective work done, or effort units. For simplicity, the price of the output is taken by the firm as given and is set equal to one.

As is quite evident, the treatment of total sales and the physical input-output aspect of production in the model is similar to its neoclassical—or Walrasian—analogue. The difference emerges when we consider the cost function. The labor argument in the production function—work effort—bears no market price, for it is labor time, not work itself, that is purchased. Hence the cost of labor—work—cannot be expressed in the firm's cost function as a market-determined hourly wage rate multiplied by the number of labor hours hired.<sup>8</sup> To express the cost function and the

---

here as well. Unlike contract theory, I assume away problem of risk aversion and issues of reputation (workers and capitalists alike have no memories).

<sup>8</sup>Marx (1976) dramatized the fact that labor itself cannot be bought and hence has no price as follows. "On the surface of bourgeois society the worker's wage appears as the price of labor, as a certain quantity of money that is paid for a certain quantity of labor" (p. 675). But "it is not labor which daily confronts the possessor of money (the capitalist, SB) on the commodity market, but rather the worker. What the worker is selling is his labor power" (p. 677). As a result, "according to the amount of actual labor supplied every day, the same... wage may represent very different prices of labor, i.e., very different sums of money paid for the same quantity of labor" (p. 683). Marx then makes it clear that the cost of a given amount of labor may vary through the extension of the length of the working day, or through an increase in the intensity of work in any given hour. "The rise in... wages may therefore be unaccompanied by any change in the price of labor, or may even be accompanied by a fall in the latter" (p. 684). Henry Ford may have understood this when he paid his workers in Detroit the unheard of sum of \$5 a day. That labor itself cannot be purchased has long been recognized outside the Marxian tradition as well. Gary Becker observed that "any enforceable contract could at best specify the hours required on a job, not the quality of the performance" (1962, p. 6). But this fact has not been given the importance it has received among Marxian economists.



production and total sales function in the same terms, a third equation is required—the labor extraction function—representing the amount of labor done per hour of labor hired as a function of the costly inputs used to elicit work from workers.<sup>9</sup>

We may write  $L$ , the total labor input, as the product of the hours of labor power hired,  $Lp$ , and the amount of work done per hour  $l^*$ , or  $L = Lpl^*$ . The amount of work done per hour is determined by the worker in response to the constraints devised by the employer, given the availability of other jobs, unemployment insurance, and the like. At this point, attention need only be given to those determinants of the worker's effort that appear as instruments from the standpoint of the employer.

The amount of work done per hour will depend upon the worker's perception of the cost of pursuing a nonwork activity, that is, of acting on the basis of any of his or her nonwork (and work-reducing) objectives. Assuming that a worker's job will be terminated if the worker's nonwork activities are detected, the expected cost of pursuing nonwork activities,  $E(n)$ , is the product of two terms: the probability that a worker's nonwork strategy will be observed by the employer,  $p^o$ , and the cost of being fired, if observed,  $w^*$ . It is assumed that  $p^o$  is positively affected by the amount of surveillance inputs (material or human) purchased per

hour of production labor hired,  $s$ , or  $p^o = p^o(s)$ , and  $p^o(0) = 0$ , and  $p_{os} > 0$  for  $s > 0$ . (Here and below subscripted functions indicate the partial derivative of the function with respect to the variable indicated by the subscript.)

Surveillance labor does not enter into the transformation of inputs into outputs, and is thus distinct from what may be termed coordination labor, which is a production input represented in the production function as a component of  $L$ . (Here I abstract from the far from trivial problem of extracting work from surveillance employees. Thus I represent surveillance services,  $s$ , as purchasable at price  $p_s$ .) The cost of an hour of labor power,  $c_{Lp}$ , is thus  $(w + p_s s)$ , and the cost of an effort unit of labor,  $c_l$ , or what Marx called the price of labor, is  $(w + p_s s)/l^*$ .

The money cost of being fired is measured by  $w^*$ , the difference between the wage offered and the worker's expected income if fired. (I assume for simplicity that the worker has no nonwage income if employed.) This latter term is simply a weighted average of  $w^c$ , the worker's nonwage income if fired and not reemployed (unemployment insurance, means-tested income support payments, and the like), and  $\hat{w}$  the expected wage in some other job, should the fired worker find employment elsewhere. It is assumed that both wages ( $w, \hat{w}$ ) exceed  $w^c$ . Thus assuming a time horizon of a single period and letting  $j$  represent the probability of finding another job (or equivalently, the fraction of the period during which the worker expects to remain unemployed), the expected income loss,  $\hat{w}^d$ , is

$$\hat{w}^d = w - [j\hat{w} + (1-j)w^c].$$

All of these wage terms, including  $w^c$ , are expressed in real units.<sup>10</sup>

A particularly simple model of the worker's response to the employer's choice of various

<sup>9</sup>Note that if labor costs did *not* depend on hours of labor hired but only on the amount of labor done, or if the relationship between hours hired and work effort performed were exogenously determined, or if the extraction of work from workers were costless, the third equation would be unnecessary. However, even the use of straight piece-rate payments will not render costs independent of the hours of labor hired unless the piece-rate workers use no inputs owned by the firm, and the determination of the number of pieces produced requires no surveillance inputs and hence is costless. But in this extreme case, there is no reason—by conventional definitions—to consider the piece-rate workers part of the firm that purchases their output, for their sole relationship to the firm is an exchange. The necessity for the third equation is thus based on assumptions no different from those used in the Coasian tradition to explain the existence of firms. The manner in which this function is developed is quite different, as we shall see, from its Coasian analogue.

<sup>10</sup>Note that because the employer clearly may directly set only nominal variables, but seeks to implement a real strategy, the general price level will enter into the employer's wage setting even in the absence of cost-of-living provisions in contracts. But I will not develop this point here.

P 3125 (3125)

combinations of surveillance and wage-loss threat results if we assume that at any moment the worker's decision is to work at a level of intensity satisfactory to the employer or not to work. The intensity of labor,  $l^*$ , then is just the percentage of time on the job during which the worker is actually working. It is assumed that the worker chooses a desired level of  $l^*$ , and then selects the moments of work and nonwork randomly. The probability that the worker will be detected not working, and hence dismissed, ( $p^d$ ), is equal to the probability of being observed at any moment ( $p^o$ ), multiplied by the probability that at that moment the worker will not be working ( $1-l^*$ ), or  $p^d = p^o(1-l^*)$ . The probability of job retention is simply  $(1-p^d)$ , setting aside reasons for job termination other than observed nonwork. Thus, for  $l^*=1$ ,  $p^d=0$ .

Let us assume for simplicity a two-period framework in which hiring occurs only at the beginning of a period and firing occurs only at the end of a period. The worker's time preference is assumed to be zero. The worker's expected income over two periods is thus the first-period's (assured) wage plus the expected wage or nonwage income for the second period:

$$\hat{y} = w + (1-p^d)w + p^d(j\hat{w} + (1-j)w^c).$$

Assuming identical workers and employers makes it reasonable to represent the worker as perceiving the alternative wage as identical to the present wage, or  $w = \hat{w}$ , and thus the expected income in the second period, if dismissed at the end of the first period would be  $w - \hat{w}^d$ , and rewriting the above expression for  $\hat{y}$ :

$$(2) \quad \hat{y} = 2w - p^d\hat{w}^d.$$

The worker's expected effort over two periods is both the effort expended in the current job, and the effort expended in the next job, should the worker be terminated and then reemployed. (Given the assumption that the worker has full information and hence nothing to learn, it is reasonable to suppose that the worker's choice concerning work effort when reemployed will be identical to the

prejob loss choice.) Thus, the expected level of effort is

$$(3) \quad \hat{l}^* = l^* + (1-p^d)l^* + p^djl^*.$$

The worker values income and, on the margin at least, finds increased work intensity displeasing.<sup>11</sup> The risk-neutral worker's response to the employer's strategy will be that which maximizes

$$(4) \quad \hat{u} = \hat{u}(\hat{y}, \hat{l}^*)$$

by equating the expected marginal disutility of effort (from equations (3) and (4)) with the expected marginal utility of income associated with an increment of effort (from equations (2) and (4)).<sup>12</sup>

Because the expected marginal income return to an increment in work will depend positively on  $\hat{w}^d$ , under quite general assumptions it can be shown that the worker's choice of  $l^*$  will be a positive function of  $\hat{w}^d$ .<sup>13</sup> By similar reasoning it can be shown that work intensity will be a positive function of  $s$ .

We may now represent the amount of work done per hour of labor power pur-

<sup>11</sup>This does not require a marginal disutility of labor (or effort). Even on the margin, the worker may enjoy the process of work, or despise it; what is essential to my argument is the assumption that the workers' objective function includes some positively valued on-the-job activities (or inactivity) that are associated with a positive opportunity cost in terms of working.

<sup>12</sup>That is, by equating  $(\partial u / \partial \hat{y})(\partial \hat{y} / \partial l^*)$  with  $-(\partial \hat{u} / \partial l^*)(\partial \hat{l}^* / \partial l^*)$ .

<sup>13</sup>Assuming the second-order conditions for the worker's utility maximization to be met, it can be shown that effort will be an increasing function of  $\hat{w}^d$  for  $\hat{w}^d \geq 0$ ,  $l^* < 1$ , and  $s > 0$ . This is because an increase in  $\hat{w}^d$  will increase  $(\partial \hat{u} / \partial \hat{y})(\partial \hat{y} / \partial l^*)$ . This follows readily from the independence of  $\partial \hat{u} / \partial \hat{y}$  from  $\hat{w}^d$ , and the fact that  $\partial \hat{y} / \partial l^* = p^d\hat{w}^d$ . Thus,  $(\partial^2 \hat{y} / \partial l^* \partial \hat{w}^d)$  must also be positive (for  $s, \hat{w}^d > 0$ ). The upward shift of  $(\partial \hat{u} / \partial \hat{y})(\partial \hat{y} / \partial l^*)$  associated with an increment in effort will necessarily result in an increase in effort as long as the disutility associated with a marginal increment in effort is not infinite. Thus the derivative of work effort with respect to the cost of job loss will be positive for positive  $\hat{w}^d$  and  $s$ . Assuming the expected marginal utility of effort is independent of  $s$ , the analogous result for  $s$  follows.

chased,  $l^*$ , as

$$(5) \quad l^* = h(s, \hat{w}^d).$$

The function  $h$ —the labor extraction function—summarizes the effects of all of the relevant preferences of the worker, as well as the worker's sense of commitment, injustice, resentment, deference, patriotism, or whatever may affect the difficulty or ease of extracting labor from labor power, or influence the efficacy of surveillance or the threat of income loss as instruments towards this objective.<sup>14</sup>

<sup>14</sup> Before bringing together the three functions—production, cost, and extraction—to consider formally the capitalist's profit-maximizing problem, it may be useful to scrutinize more carefully the nature of the extraction problem. Is this not just another case of the economics of lemons, in which the employer must pay some costs to find out which workers will work hard (or well) and which will not? While some of the results are similar, not all are, and the mechanisms are quite different. The problem for the employer is not to find out what the worker *is*, but to find out what the worker *does*. To see that this is the case, the extreme assumption is made that the employer may know at zero cost the workers' skills and personality characteristics relevant to work motivation and capacities, including exact knowledge of the determinants of the typical (and therefore every) worker's work effort. One of the determinants of work effort is the threat of job loss and hence the level of surveillance. The employer, by these assumptions, knows exactly how much work each worker will do on the average once the employer has selected the level of surveillance and the wage (given external wages, unemployment probabilities and unemployment insurance). At a given moment, however, the employer does not know what the worker is doing, unless the worker is being observed at that moment. And unless the worker is observed not working up to standard, it would not be rational for the employer to fire him or her, for this would convince the remaining workers that the probability of job loss did not depend on work effort, and would thus lower the efficacy of the surveillance inputs. Note that by firing the worker the employer does not eliminate a "bad worker" in favor of a chance at getting a "better worker" from the unemployment pool, for all workers are identical. The purpose of firing the nonworking worker is to convince workers that the surveillance system is effective, and that firing is related to low work effort. In other words, without firings or with firings not based on observed low work effort, the  $h$  function would shift adversely from the standpoint of the employer. Strictly speaking, then, the cost of surveillance is not an information cost at all (or at least a very peculiar one) as surveillance will affect increases in effort (over

It is assumed that the employer knows the  $h$  function of each worker, and that each is identical, thus allowing one to argue in terms of a representative worker. Further, on the basis of the reasoning above, for both  $s$  and  $\hat{w}^d$  positive and  $l^* < 1$ ,  $h_s$ , and  $h_{\hat{w}^d}$  are positive, and  $h_{s\hat{w}^d}$  is also positive.<sup>15</sup>

Letting  $p_x$  represent a vector of prices of nonlabor inputs, the problem for the employer is now to maximize

$$(6) \quad R = f(X, L) - p_x X - (w + p_s s) L p,$$

subject to

$$(7) \quad L = l^* L p = h(s, \hat{w}^d) L p,$$

or to maximize

$$(8) \quad R = f[X, h(s, \hat{w}^d) L p] \\ - p_x X - (w + p_s s) L p.$$

Because it has been assumed for the moment that the nonlabor inputs  $X$  do not affect the labor extraction process, the production function and the extraction function

some range) even if the "surveillors" do not pass the information along to the employer, as long as the workers believe that the probability that a nonwork strategy will be detected is a positive function of the level of surveillance. But if employers know exactly how much work each worker will do once the wage and level of surveillance is selected, would it not be optimal to pay workers according to the amount of work done? It might. But this in no way would affect the results below, for the firm's costs will still depend on the number of hours hired (because surveillance  $s$  is proportional to hours of labor engaged, not the amount of work done and because workers use inputs owned by the firm). And as long as costs are not independent of the number of hours hired, employers will not be indifferent to how hard each particular worker works. (We will see below that the limiting case of no surveillance inputs cannot be optimal. It is, of course, possible to devise combinations of incentive pay and surveillance such that costs would be independent of hours hired. But it would be quite accidental if that scheme coincided with the optimal incentive structure, given workers' preferences and other relevant information.)

<sup>15</sup> More formally, because the derivative of expected income with respect to work intensity is simply  $\hat{w}^d p^d$ , the effect of an increase in  $\hat{w}^d$  on the workers optimal effort level will depend positively on the level of  $s$ , and conversely.

(equations (1) and (5)) are separable, and the employer's maximizing problem may be solved sequentially. The first problem for the employer, and the one that interests us here, is to minimize the cost of a unit of work done, or

$$(9) \quad \min c_l = (w + p_s s) / h(s, \hat{w}^d).$$

Having solved this problem, its solution,  $c_l^0$ , can then be considered the minimum cost of a unit of labor and entered into the employer's new maximand

$$(8') \quad R = f(X, L) - c_l^0 L - p_x X.$$

Assuming, for the moment, an interior solution, and noticing that the marginal cost of a unit increase in  $\hat{w}^d$  is one by definition, minimizing (9) requires that

$$(10) \quad h_{\hat{w}^d} = h(s, \hat{w}^d) / (w + p_s s) = h_s / p_s,$$

or that the average effort per dollar of wage and surveillance cost equal the marginal effort per dollar increase in either wage cost or surveillance cost. Analogously the profit-maximizing employer's strategy must satisfy the condition

$$(10') \quad p_s = h_s / h_{\hat{w}^d},$$

or the price of surveillance must be equal to the "marginal rate of substitution" between income loss if fired and probability of detection in the labor-extraction function (5).

We may represent this graphically as in Figure 1. The isocost function is a locus of equally costly employer strategies. Because the cost to the employer of a unit increment in  $\hat{w}^d$  is one by definition, the slope of the isocost function can be seen to be  $-p_s$ . The isowork function, derived from the labor extraction function (5), is one of a family of loci of equally effective employer strategies: points describing an equal extraction of labor from a given number of hours of labor power hired. Its slope is  $-h_s / h_{\hat{w}^d}$ . The expansion path is the locus of all possibly profit-maximizing strategies, namely, those satisfying (10'). Some point on the expansion path, say point  $a$ , minimizes the cost of a unit of labor

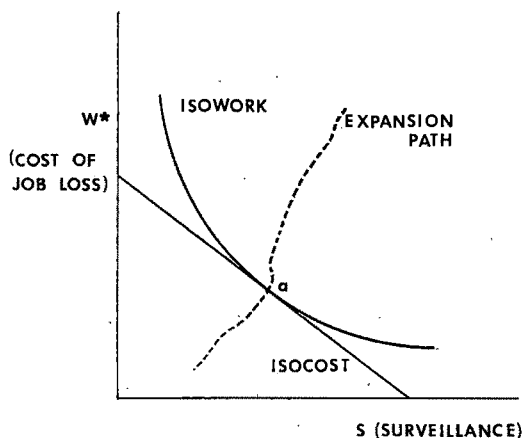


FIGURE 1

and is therefore the solution to (9) and is the profit-maximizing strategy. (It cannot be read directly off the figure.)

I now use this model to demonstrate the three substantive propositions with which I began.

## II. The Reserve Army of the Unemployed

The more or less permanent existence of involuntary unemployment is central not only to the Marxian critique of capitalist society, but to the analytical underpinnings of its theory of profit (or of surplus value) as well. Because profits in the Marxian model are not a return to a scarce input but are simply a deduction from total output made possible by capital's power over labor, a complete Marxian model must provide a compelling account of how this power is perpetuated in an economically competitive and politically liberal environment. The basis of this account is the asymmetry between two forms of competition: that among capitalists in selling their outputs and that among workers in seeking employment. Because profit is not the return to a scarce input, in the absence of such an asymmetry, there would be no reason why price competition among capitalists would not drive the profit rate to zero.

The necessary asymmetry is based on the permanent existence of involuntary unemployment, or what Marx termed the reserve

army of the unemployed.<sup>16</sup> The effect of involuntary unemployment is to render labor power nonscarce and hence incapable of claiming the whole product (net of depreciation) through the normal process of competitive price and wage determination. The puzzle is then no longer why profits are not competed away, but why does a nonscarce input, labor power, receive any competitive remuneration at all. The capital-labor distributional conflict thus appears as one taking place between and among two sets of actors, none of which exercise their claims on the product on the basis of a competitively determined return to scarcity in the usual general equilibrium sense.

The Marxian solution to this puzzle is to not reject the competitive assumptions underlying the general equilibrium model, but to pose a distinct theory of the long-term determination of wages and effort in which the former varies negatively and the latter positively with the level of unemployment.<sup>17</sup> Only *involuntary* unemployment will affect the bargaining power of capital and labor; hence the centrality of involuntary unemployment to the Marxian theory of the capitalist economy.

On what basis can involuntary unemployment be represented as a general—rather than ephemeral—characteristic of the capitalist economy? The endogenous perpetuation of the reserve army of the unemployed could be assured by a variety of mechanisms: for example, an infinitely elastic supply of labor from other countries or from declining domestic noncapitalist economic systems, such as household production, or rapid structural and technical change accom-

panied by downwardly sticky wages.<sup>18</sup> The above model of the extraction of labor from labor power points to another possibility, and one more consistent with competitive assumptions, namely, that the labor market does not clear in equilibrium. Put somewhat differently, excess supply in labor markets does not imply a competitive response of wage reductions.

By equilibrium in the labor market, I mean a level of wages, employment, and labor intensity that none of the agents would have both the motivation and the ability to alter. A non-clearing-labor-market equilibrium requires that profit-maximizing employers offer workers a wage and surveillance package such that, given the levels of work effort that workers will choose to expend under the package offered, workers are not indifferent between working and being unemployed. This is, of course, tantamount to saying that a profit-maximizing employer would refuse the offer by a currently unemployed worker to work as hard as the current work force for a wage less than the current wage. We shall see why this counterintuitive result may quite generally occur.

It is clear then that a market-clearing wage would imply that in our model the cost of job loss is zero, for if the cost is not zero the worker cannot be indifferent between employment and unemployment. Under what conditions could a wage-surveillance package that rendered the worker indifferent between employment and nonemployment be a profit-maximizing strategy for the individual employer, and hence a possible equilibrium? Or, in terms of Figure 1, could an optimal strategy lie on our horizontal axis, indicating a zero income loss associated with being fired? Because  $h_s \rightarrow 0$  as  $\hat{w}^d \rightarrow 0$ , and analogously  $h_{\hat{w}^d} \rightarrow 0$  as  $s \rightarrow 0$ , the expansion path for any  $p_s > 0$  will lie entirely within the range of positive values of  $s$  and  $\hat{w}^d$ . As long as the employer has hired some surveillance inputs, a market-clearing wage ( $\hat{w}^d = 0$ ) can-

<sup>16</sup> Marx (1976): "...relative surplus population (i.e., unemployment, SB) is therefore the background against which the demand and supply of labor does its work" (p. 792). And, "The pressure of unemployment compels those who are employed to furnish more labor and therefore makes the supply of labor to a certain extent independent of the supply of workers. The movement of the law of supply and demand on this basis completes the domination of capital" (p. 793).

<sup>17</sup> The macroeconomic and general equilibrium characteristics of this solution are the subject of two of my other papers (1983a,b).

<sup>18</sup> If the supply of labor hours is infinitely elastic at a given wage, those who are not employed cannot be said to be involuntarily unemployed strictly speaking, as they are unwilling to offer any labor time at a lower wage.

not be optimal (because  $h_s = 0$  for  $\hat{w}^d = 0$ ). The critical role of the cost of surveillance is here clearly indicated, for with  $p_s = 0$  the isocost functions in Figure 1 would be horizontal: (free) surveillance would be substituted for (costly) job loss threat and the cost minimum would occur at  $\hat{w}^d = 0$ , a result consistent with the traditional market-clearing equilibrium.

But what of the "no income loss, no surveillance strategy" represented by the origin in Figure 1? In order for this strategy to be optimal, it would have to be the case that

$$(11) \quad h(0,0)/w > h(s, \hat{w}^d)/(w + p_s s)$$

for all possible levels of  $s$  and  $w$ . In this case, surveillance and job loss threats are sufficiently ineffective or costly to prohibit their use at any level. But this implies that, even when it is possible for the employer to exercise power over the worker, it is not profitable to do so. But this could only be true if there were no conflict of interest between the worker and the employer. In this case, employer and workers have a "conflict of interest" only in the socially irrelevant sense that sunbathers and drought-stricken farmers have a conflict of interest (barring the possibility of rainmaking).

This result does not depend on the manner in which the probability of reemployment ( $j$ ) is determined. Assume for the moment that the government committed itself to achieving full employment, either through fiscal and monetary policy, or simply by guaranteeing any unemployed worker a job at the going wage. With  $j=1$  the employer might either set  $\hat{w}^d > 0$  by offering a wage higher than other employers, or set  $\hat{w}^d = 0$ . The former is inconsistent with equilibrium. This can be readily seen by rewriting the cost of being fired as an equilibrium condition (with  $w = \hat{w}$ ) or  $\hat{w}^d = (1-j)(w - w^c)$ . By the logic of the previous paragraph, the latter is inconsistent with the assumed conflict of interest between worker and capitalist.

Let us summarize these results. Given a positive cost of surveillance and a conflict of interest between employer and worker over work effort, the wage rate offered by the

competitive profit-maximizing employer will exceed the worker's next best alternative. This is possible in general only if the probability of reemployment is less than one. Therefore, labor market competition cannot clear the labor market. Correspondingly, market clearing—the absence of involuntary unemployment—implies labor market disequilibrium.<sup>19</sup>

Other than ruling out market clearing as a possible labor market equilibrium, this model bears no direct implications concerning the determination of the general level of unemployment or the probability of reemployment. But it does provide a microeconomic foundation consistent with Kalecki's suggestion that sustained full employment and the long-run survival of capitalist enterprise may be inconsistent. Indeed, given a conflict of interest between employer and worker, labor market clearing implies either escalating wage increases, or a reduction in work effort to those levels chosen by workers. Particularly in an open economy, neither result would likely be conducive to investment levels capable of sustaining full employment (but to pursue this argument we would have to go considerably beyond the microeconomic confines of this paper).<sup>20</sup>

These results would be modified, of course, if employers were assumed to have not pro-

<sup>19</sup>This result is similar to that produced—with somewhat different models—by Calvo, B. Curtis Eaton and William White (1982), James Malcomson (1981), Hajime Miyazaki (1981), Tekashi Negishi (1979), Solow (1980), Carl Shapiro and Joseph Stiglitz (1984) and others. In all of the above, actual amount of work done is directly or indirectly a positive function of the wage rate. Miyazaki focuses on the problem of worker free riding against other workers in a work group. Eaton and White focus on "trust jobs." Malcomson assumes "at least two types of individuals with different productivities who cannot be discriminated perfectly by observation at work" (p. 865). Negishi and Solow both base their models on problems of worker morale and "affront" (Negishi, p. 114). Closest in spirit to my model (though lacking the surveillance element) is Calvo, who, however, while demonstrating the possibility of nonclearing equilibria, assumes an interior solution to a problem analogous to the minimization of (6), thus eliminating the market-clearing equilibrium by assumption.

<sup>20</sup>I develop this argument in my 1981, 1983a, b papers. See also Gintis and Tsuneo Ishikawa (1983).

hibitively expensive ways of imposing effective sanctions on workers even in the absence of involuntary unemployment. The extent to which such alternative sanctions are feasible and effective is in part an empirical issue that cannot be resolved here. For whatever reason, the practical import of most of the alternatives to the threat of involuntary unemployment appears to be quite limited in the U.S. economy.<sup>21</sup>

### III. Capitalist Technology

Central to the Marxian critique of capitalist society is the idea that the competitive pursuit of profits requires employers to organize the production process so as to maintain their power over workers, and that at least some of the boredom, fragmentation, and other undesirable aspects of the work experience may be attributed to this fact and not to the requirements of technical rationality. According to this view, the prevailing organization of production—including the technologies in use—cannot be derived solely from an interaction of exogenously given technical possibilities and worker and consumer preferences for goods, leisure, and various kinds of work environment, but rather reflect the class interest of capital as well. Hence the expression “capitalist technology.”

To suggest that technology may be an instrument of class conflict does not mean, of course, that employers may select technologies without regard to the competitive requirements of cost minimization. Nor does it require that capitalists collude in their choice of production methods or in the development of future technologies. Rather, the concept of capitalist technology is based on the proposition that cost minimization by competitive employers implies the selection of profitable but inefficient technologies even in the absence of market failures arising from collusion, externalities, extended time horizons, and the like.

I will say that the capitalist has chosen an inefficient technology when there exists some other method of production that, per unit of output, uses less of at least some input and not more of any. The logic of the concept of capitalist technology is that a technology that is inefficient in the above sense may nonetheless be cost minimizing if it allows the capitalist to lower the cost of some input. This is possible in the Marxian model because the firm is not a price taker with respect to the price of labor, but rather may alter this cost through the selection of various labor extraction strategies. The most obvious case of this is the adoption of machine-paced production as a means of

<sup>21</sup>If workers could instantaneously find alternative employment, but nonetheless bore significant costs of job changing—either through moving costs, training costs not borne by their new employer, employment bonds, or job entry fees that are forfeited upon job loss, a tax levied by the government on job changers, or through any other means, or if on-the-job nonwork activities were treated as a criminal offense subject to fines or imprisonment, the attainment of full employment could not be ruled out on theoretical grounds. While possible substitutes (or complements) to the threat of unemployment are thus readily imaginable, their actual or potential relevance to the problem of getting workers to work may be questioned. First, to replace the threat of unemployment, the costs imposed must be quite substantial, considerably more than reasonable moving or training costs, and in excess of what most workers can readily borrow for payment of an employment bond. Juliet Schor and I (1983) estimate that in 1983, for example, the mean cost of job loss (roughly an after-tax estimate of  $\bar{w}^d$ ) was about one-half the mean after-tax annual income of a fully employed production worker. (This is a low estimate, as it abstracts from the costs associated with the loss of job seniority.) Moreover, the variance among individuals of the expected cost of job loss is probably quite large, due to the high variance of unemployment duration, suggesting that if we were to drop the unrealistic assumption that workers are risk neutral, the certainty equivalent of the cost of job loss might be considerably greater than Schor's and my estimates. Consideration of the social or psychological costs of unemployment—even with a generous accounting of the joys of free time—would further augment the estimate of the costs of job loss. Second, the imposition of these alternative sanctions by either employers or through the government may involve private or social enforcement costs, or other welfare losses sufficiently large to inhibit their use. Third, some otherwise promising methods of eliciting work effort other than the threat of unemployment may be considered to be socially unacceptable or politically infeasible. Even assuming that effective alternative sanctions were feasible would only modify rather than nullify my results unless these alternatives were so cost effective as to totally eclipse the expedient of paying workers more than their supply price.

increasing the intensity of labor.<sup>22</sup> In this case, costs may be lowered not only by producing more with the same inputs, but by extracting more of one of the inputs—labor—for the same price, and thus lowering the unit cost of labor. Machine-paced production may of course also be efficient. But it is simple to show that it *need not* be efficient in order to be adopted.

Capital goods may be considered to be capable of joint production, simultaneously contributing to the marketed output of the firm and producing or contributing to the acquisition of information on the work performance of the workforce. The assembly line, and even factory production itself (in contradistinction to more decentralized production methods), as well as modern information-processing systems are important cases of surveillance information-producing technologies.

The implications for efficient technical choice may be readily seen by modifying the labor extraction function to take account of this form of joint production. We now have

$$(5') \quad l^* = h[p^o(s, x), \hat{w}^d],$$

where  $x$  is the vector of inputs (per labor hour) of production equipment and intermediate goods, and  $p^o(s, x)$  is the worker's expected probability that a nonwork strategy will be detected. For some  $x$  we have  $p_{ox} > 0$  and hence  $h_x > 0$ : given the cost of job loss ( $\hat{w}^d$ ) and the level of (pure) surveillance inputs ( $s$ ), the use of larger amounts of some input in the production process will increase the amount of work done per hour by increasing the probability that a nonwork strategy will be detected, thus increasing the worker's expected cost of pursuing a nonwork strategy.

It can be seen in this case that even if all relative goods prices were optimal (in the sense that they accurately reflected relative scarcities), the familiar conditions for efficient technical choice (i.e.,  $f_x = p_x$ ) would be violated. For it will now be the case that the

profit-maximizing employer will maximize profits by observing the following condition:

$$(12) \quad f_x + f_l h_x = p_x.$$

The second term on the left-hand side reflects the contribution of a marginal increment in  $x$  to production via its contribution to the extraction of labor from labor power. (It is redundant to observe that under these conditions the relative general equilibrium prices would also not be optimal.)

The implication of this point is that a competitive profit-maximizing capitalist could choose a technology using more of both  $x$  and  $l^*$  per unit of output. This may be readily seen by noting that the isocost function slope is

$$(13) \quad dl^*/dx = -(p_x + l^* c_{lx})/c_l,$$

where  $c_{lx}$ , the derivative of the cost of a unit of effort with respect to  $x$ , is negative, and hence the numerator is not necessarily negative. Thus the isocost function may be positively sloped, leading to the possibility that cost minimization may result in the choice of an inefficient technology, namely in the rejection of a technology using less of both  $l^*$  and  $x$  per unit of output.

It might be thought that this demonstration implies that the need for surveillance inputs is somehow illegitimate and should be abstracted from in consideration of efficiency. Indeed, as we shall see in the penultimate section, the assertion that the class structure of capitalism induces a particularly high level of work resistance and hence promotes the extensive use of surveillance inputs differentiates the Marxian from the neo-Hobbesian view. But the above argument involves neither abstracting from surveillance inputs, nor considering surveillance to be a kind of false need induced through an endogenously generated disutility of labor.

Quite the contrary, pure surveillance inputs  $s$ , with an *exogenously* determined labor extraction function, provide a particularly clear case of the above argument. Consider the indicated isowork function in Figure 1 as representing an amount of work effort capable of producing one unit of output. Starting at point  $a$ , were the firm to move along this

<sup>22</sup> Edwards (1979) refers to this as "technical control" in contradistinction to "bureaucratic control" or "simple control" of the production process.



isowork locus by raising wages and cutting surveillance inputs, the cost of labor would rise and hence the profit rate would fall, but output per unit of input would rise ( $I^*$  remaining constant and  $s$  falling). This result arises because there is a tradeoff between surveillance and the wage rate in the labor extraction function, and while surveillance inputs are resource-using, the wage rate is not; hence raising wages and lowering surveillance may be efficient but not profitable. Thus cost minimization and efficiency do not coincide: the tradeoff in this case is not efficiency vs. equity, but efficiency vs. profitability.

#### IV. Divide and Rule

Central to recent Marxian research on racial and sexual discrimination, segmented labor markets, and internal labor markets is the proposition that divisions among workers may be in the interest of employers, and further that it may be in the interest of competitive noncolluding employers to discriminate among workers on the basis of ascriptive characteristics unrelated to the individual worker's ability or willingness to contribute to the production process.<sup>23</sup>

Reich, Roemer, Gintis, and others have recently proposed coherent models of discriminating competitive capitalists. The present model of the extraction of labor from labor power may be extended in a very simple way to capture the logic of these contributions.<sup>24</sup>

<sup>23</sup> This view may be distinguished from that which maintains that the cost-minimizing process renders discrimination unprofitable to the individual employer, however beneficial it might be to the employer's class as a whole, and hence that discrimination is primarily an ideological or political phenomenon whose perpetuation is explained by inertia, ignorance, or by the collective action (in the media, schools, state, or elsewhere) of those who benefit from it.

<sup>24</sup> See Reich and the previously cited references to Roemer and Gintis. This model differs somewhat from those cited in stressing the costly nature of surveillance and the cost of job loss rather than bargaining strength based on worker unity. All of the Marxian models differ from the search theory approach to the stability of discrimination in a competitive environment in that the employer is assumed to know all of the relevant worker characteristics.

We say that an employer discriminates when he or she makes different wage-surveillance offers to workers of differing ascriptive characters (race, sex, age) who are otherwise identical with respect to their productive capacities and proclivities, that is, given that we have assumed that labor services are homogeneous, identical with respect to their labor extraction functions,  $h$ . I now introduce the possibility that workers may cooperate either to render surveillance more difficult or otherwise more expensive (for example by refusing to offer information on the work or nonwork activities of fellow workers), or to reduce or withdraw labor services should the employer treat a fellow worker in a manner thought to be unjust or simply contrary to the interests of other workers. Labor services may be withdrawn either through a reduction in work effort (an outward shift in the  $h$  function), or in an extreme case through a strike.

The extent of worker cooperation, including the possibility of forming institutions such as unions, varies positively with the extent of worker unity,  $u$ . Worker unity will depend on general social conditions external to the firm, but it will also be influenced by the firm's hiring and pay policies. Where a uniform wage surveillance package is offered to all workers, for example, opportunities for joint negotiations concerning wage and working conditions will be enhanced, and divisive sentiments such as envy and invidious distinction attenuated. With distinct pay and surveillance packages offered to different workers—particularly to groups of workers predominantly composed of individuals of different race, sex, age, and other characteristics—employers are more likely to be able to bargain separately with each group to foster competition, envy, or even hostility among the distinct groups, and thus to discourage unity. For simplicity we say that unity,  $u$ , will be a negative function of a measure of wage inequality of the workforce of the firm,  $v$ .<sup>25</sup>

<sup>25</sup> Because  $u$  cannot readily be measured, this behavioral assumption cannot easily be tested. But it is strongly supported by the relevant works in labor economic and labor history. See Reich, and Edwards, Gordon, and Reich and the works cited therein.

I make an additional assumption, not necessary to my result but one which will enrich the model somewhat: let us now assume that there are some costs to the employer of replacing the worker (firm-specific training, or other), and that, for this reason, when a worker is detected pursuing a nonwork strategy, the employer may choose not to terminate the worker's employment.

In my expanded model, then, the expected cost to the worker of pursuing a nonwork strategy is

$$(14) \quad E(n) = p^o p^t \hat{w}^d,$$

where, as before  $\hat{w}^d$  is the cost of job loss,  $p^o$  is the probability of being detected should the worker pursue a nonwork activity, and  $p^t$ , previously assumed to be unity, is now the variable probability of being terminated, if detected. By the above argument,

$$\begin{aligned} p^o &= p^o(s, x, u) \quad \text{with } p_{ou} < 0; \\ p^t &= p^t(u) \quad \text{with } p_{tu} < 0. \end{aligned}$$

The labor extraction function thus becomes

$$(5'') \quad l^* = h[p^o(s, x, u), p^t(u), \hat{w}^d]$$

in which the derivative of  $l^*$  with respect to  $u$  is negative, taking account of the effects of unity on both the probability of detection and the probability of termination.

Under what conditions will the employer described in this model choose to discriminate? Assume that there are two "types" of worker, type  $i$  and type  $j$ . Why would the employer pay them different wages? It is clear at once that if the wage rates prevailing in the rest of the economy are different, or if the probability of reemployment or access to unemployment insurance is different, the optimal wage offers  $w_i$  and  $w_j$  will differ. Thus given differing external conditions, the firm will choose to offer differing wages to each type of worker. But it will be clear that the cost of a unit labor from one type of worker is less than the other, or  $c_{ij} < c_{ii}$  (assuming that type  $i$  workers are favored by higher wages and/or reemployment probabilities or access to unemployment insurance in the remainder of the economy). So the question

arises, why would the employer choose to employ any of type  $i$ ?

Assume that the employer hired no type  $i$  workers. In this case, there would be no wage inequality among the workforce ( $v = o$ ). Hiring some type  $i$  workers will yield a positive  $v$ , thus increasing  $l^*$  and possibly lowering the average cost of labor for the firm as a whole,  $c_j$ . By the same reasoning, it could be in the interest of the employer to offer type  $i$  and type  $j$  workers different wages, even if in the rest of the economy they were treated perfectly equally. Moreover, given the existence of involuntary unemployment, such a strategy would not be rendered infeasible by the labor supply choices of the group which was offered the lower wage.

A related but distinct argument for paying identical workers different wages may also be offered, if the model is extended to more than one time period. Assume initially that all workers are paid the same wage. An employer could then offer a prospective worker a two-period wage package with a low first-period and high second-period wage. The difference in the first-period wages under the equal wage and the stepped-wage package may be considered an employment bond paid by the worker to the employer which will be returned to the worker in the form of higher second-period wages, unless, of course, the worker is fired in the interim. Let the wage cost to the firm of the two packages be the same, assuming the firm intends to make good its second-period offer, and expects the worker to neither quit, nor be fired. The "less now, more later" offer will elicit more work from the worker, however, because once it is accepted and work under its terms has commenced, the cost of job loss under the terms of that package is greater, because the worker has already performed some low wage labor and has an increasingly advantageous balance of high wage labor to look forward to should he or she retain the job. In a regime of generalized stepped-wage offers such as the primary labor market in the United States, the costs of failing to cash in on later-period high wages can be considerable.

The worker may not accept the stepped-wage offer, of course, if he or she believes that the probability of getting arbitrarily fired

at the end of the first period is high. But should the worker accept the stepped-wage offer, the firm will have affected a reduction in its cost of labor  $c_l$ . As in the case of discrimination above, the fact that jobs are rationed will allow the firm to recruit labor using the less attractive stepped-wage offer.

Thus long-term contracts and internal labor markets—promotion ladders according to job tenure and unrelated to skill—may be a method of increasing the cost of job loss to the worker without increasing the wage bill, and hence an effective means of reducing the cost of labor (in effort units).<sup>26</sup>

The above explains why identical workers may be paid differently. It does not explain why discrimination exists, or why type  $i$  workers tend to be white, male, and neither very young nor very old. But it does present one possible argument for the reproducibility of discrimination and internal labor markets in a competitive capitalist economy.

#### V. Neo-Hobbesian and Marxian Models

It may well be objected that while the labor extraction model provides an internally consistent analysis of involuntary unemployment, inefficient technical choice, and discrimination in a competitive equilibrium, any negative normative connotations would be misplaced, for these undesirable outcomes might be intrinsic to *any* system of production, irrespective of the social structure in

which it is embedded. Indeed this is precisely the implication of what I have termed the neo-Hobbesian models of the production process.

Malfeasance is to the neo-Hobbesian models what class conflict is to Marxian models. The key difference between the two is this: malfeasance is a universal human proclivity—in this case based on the inherent nature of work as a disutility. By contrast, class conflict in the labor process of a capitalist economy is the result of a specific and mutable set of social institutions; the conflict over work intensity being *at least in part* the consequence of the particular organization of work and the resulting alienated nature of labor.

Samuelson's statement cited at the outset—while based on a Walrasian model—reflects the spirit of the neo-Hobbesian models as well, for it is consistent with the view that the form of the class relationship imparts nothing of importance to the production process.

Can the neo-Hobbesian position be sustained? Can the Marxian problem—class conflict over the extraction of labor from labor power—be reduced to the general problem of malfeasance? Differing ideological connotations aside, is the extraction of labor from labor power simply another way of addressing the universal problem of "shirking"?

Concern with the general problem of reconciling individual self-interest and collective rationality is hardly new, dating back at least to Hobbes. That the regulation of self-interest through the market provided a solution to the Hobbesian problem was suggested metaphorically by Mandeville during the eighteenth century and developed fully by Walras and by twentieth-century welfare economists. If all economic interactions are contractual exchanges, the conflict of self-interest and collective rationality is capable of resolution, or at least substantial attenuation.

But, as economists of all persuasions now recognize, not all economic interactions are exchanges. Coase's conception of the firm, as a command economy of nonexchange relations, is a necessary but possibly troublesome addition to any analysis of a specifically capitalist economy characterized by an

<sup>26</sup>From quite different perspectives, a similar argument has been suggested by Edwards and by Lazear. The argument is quite distinct, however, from models based on search theory and screening costs, in which the employer has an interest in retaining the worker (because of hiring costs). See, for example, Okun. The post-World War II emergence of long-term contracts and internal labor markets as characteristic of a major segment of the U.S. economy may be attributable in part to their labor extraction cost-saving aspect, to the historically low rates of unemployment in the postwar period, and to the apparent decline in the cost of job loss associated with a spell of unemployment. Further, as Lazear has pointed out, the labor extraction advantages of long-term stepped-wage offers may help explain the otherwise anomalous phenomenon of returns to job tenure significantly in excess of any empirically compelling estimates of productivity enhancement through generalized on-the-job learning. See James Medoff and Katherine Abraham (1980).

employment relation. Strikingly, the Coasian view of the capitalist economy as a multiplicity of mini-command economies operating in a sea of market exchanges is radically different from the Walrasian foundations of welfare economics, and superficially indistinguishable from the Marxian view.

The question obviously arises, then, as to the compatibility of the Coasian insight (command) and the Mandevillian solution to the Hobbesian problem (markets). Are the command relations of the firm a rational solution to the problem of the coordination of individual and group rationality? Or are they, in some sense, a market failure attributable to the successful pursuit of the interests of those who command the firm? This is the central issue dividing the neo-Hobbesian from the Marxian analysis.

Coase, basing his concept of the firm on the notion that command relations supercede market relations when the transactions costs of markets exceed the analogous costs of command and nonmarket coordination, initiated a literature which affirmed the efficiency of the hierarchical structure of the firm.<sup>27</sup> Because malfeasance is no more than an expression of the natural self-interestedness of human beings, the cost of policing malfeasance cannot be considered evidence of a failure of markets. The logic of this position can be illustrated within the terms of the Marxian model.

Let us make the (neo-Hobbesian) assumption that the labor extraction function is given by human nature. People's attitude towards work—broadly, the disutility of labor—is unrelated to the social institutions that govern the process of work. In this case, the extraction function must be considered to be exogenous, not only to the firm but to the society as a whole. Hence the various employer strategies and their results must be considered to be little more than a consequence of the (possibly lamentable but ineradicable) human tendency to avoid work. A society might nonetheless choose to discourage discrimination, to minimize involun-

tary unemployment, or to discourage the use of surveillance equipment or personnel, but they would do so only at the cost of choosing to permit a higher level of what the neo-Hobbesian literature terms free riding or shirking, and consequently a lower average level of output per hour of labor.

But the assumptions required to sustain the neo-Hobbesian view are exceptionally restrictive and implausible. We have seen in the analysis of capitalist technology that even with an exogenously given labor extraction function, the choice of technology—including the level of surveillance—which is profit maximizing will not in general be efficient: it generally will be dominated by some other less profitable and less surveillance-intensive combination of inputs.<sup>28</sup>

Perhaps more fundamentally, the assumption of an exogenous extraction function appears to be quite arbitrary. If the organization of the work process and the principles determining the distribution of the net revenues arising therefrom influence workers' attitudes towards work and hence are among the determinants of the extraction function, the neo-Hobbesian conclusions are considerably altered. In this case, there may exist some alternative set of arrangements in which a bargain could be struck in which at least one of the participants was better off and none worse off. A possible argument may be illustrated. Rewrite the labor extraction func-

<sup>27</sup>The recent literature was initiated by Alchian and Demsetz.

<sup>28</sup>Because the efficient (less surveillance-intensive) technology is less profitable it might be objected that while the neo-Hobbesian position is faulty on static efficiency grounds, a dynamic efficiency perspective, taking account of optimal levels of investment and the relationship of profits to investment, would salvage their view. But this is not the case unless it is also assumed that the current levels of investment are at or below the optimal level and further (and dubiously) that a reduction in the profit rate is necessarily associated with a decline in investment. To the extent that capitalists consume rather than invest their profits (or invest them in other economies), a decline in the profit rate does not require a reduction in the level of investment, even if the economy is operating at the level of potential output. Of course, given the institutions that define the capitalist economy, such an effect is likely to result, but it is hardly reasonable to take as given the institutions which are themselves under evaluation.

tion as

$$(5''') \quad l^* = h(i, s, \hat{w}^d, u, x),$$

where  $i$  is a vector reflecting the general institutional environment. If it could be shown that in an environment which workers perceived to be more fair, or more consistent with their self-respect, for example, they would choose to expend more effort for any given employer strategy, then it is a simple matter to demonstrate that the initial outputs could be produced with unchanged levels of labor effort in production and using less surveillance labor.<sup>29</sup> In Figure 1, the transformed institutional environment (the change in  $i$ ) would be reflected in an inward shift in the isowork loci such that the initial amount of work could be extracted with a reduced  $s$ .<sup>30</sup> The newly released surveillance labor could then be employed producing goods representing a net addition to the total product, achieved without increasing total labor hours worked and/or workers' efforts per hour.

The above argument draws directly on the third basic characteristic of the production process in the Marxian model, the joint production of commodities and workers or the endogenous nature of workers' preferences. The attitude towards work is not, according to this principle, simply a manifestation of human nature, but in part the result of the social institutions in which the production process takes place.

In the production of workers, of course, other institutions—schools, the family, political organizations, and the like—assume a critical importance. The structure of these institutions is, however, strongly albeit indirectly influenced by the structure of the production process.<sup>31</sup> Moreover, the structure of

the production process itself undoubtedly has direct effects on attitudes towards work. A more democratic structure of decision making and a more egalitarian distribution of the firm's net revenues, for example, might both reduce the incentive to pursue nonwork activities and heighten the cost of so doing by enlisting fellow workers as more ardent enforcers of the pace of work, or more willing cooperators with the surveillance system.<sup>32</sup>

The neo-Hobbesian's normative position thus seems dubious on two grounds: the discrepancy between profitability and efficiency, and the endogeneity of the labor extraction function. If the social nature of the labor extraction function is conceded and, further, if the feasibility of forms of social structure and work organization conducive of lower levels of work resistance or higher levels of work motivation is accepted,

---

William Lazonick (1978, 1981). Lazonick concluded, "Hence it can be argued that not only the institutional transformation of the capitalist enterprise but also, and perhaps more fundamentally, the institutional transformation of the larger society was required to stabilize the capital labor relation in the mass production industries" (1981, p. 36).

<sup>32</sup>Why are the potential gains to such an alternative form of work organization not sufficient to bring such worker-based enterprises into being and to assure their success in the competitive struggle with more hierarchically structured capitalist firms? If workers' attitudes toward work were determined solely and instantaneously by the work environment in which they worked, and if credit were readily available on terms no worse than those available to capitalist firms, any group of workers could form a co-op and reap the benefits of lessened surveillance. Both assumptions are highly questionable. To the extent that attitudes toward work are determined by an entire nexus of social institutions which change slowly, the opportunities for the atomistic movement towards a less socially irrational form of production are quite limited. Perhaps more important, because workers' own assets are not extensive, their access to credit is limited or costly by comparison to that enjoyed by the owners of firms. (It matters little for the issues treated here whether the different terms of credit available to capitalists and workers reflect rational profit-maximizing behavior by lenders or an imperfection in the credit market.) And it might be added that, perhaps for some of the reasons outlined in this paper, and despite the obstacles outlined in this note, the last decade has witnessed a substantial growth of workers' co-ops and worker-managed firms in the United States.

<sup>29</sup>There seems to be considerable evidence that this is the case. See, for example, Raymond Katzell et al. (1975).

<sup>30</sup>A simple reduction of  $s$  would not be optimal, of course, but this is immaterial to my argument.

<sup>31</sup>The influence is mutual, of course, schools and families influencing the structure of production as well as conversely. See, from very different perspectives, my book with Gintis (1976), Melvin Kohn (1969), and

or, if the possible nonoptimality of the competitively determined profit rate is admitted, the command relationships within the firm and the associated patterns of involuntary unemployment, technical choice, and discrimination must be viewed as market failures rather than simply as unavoidable transactions costs. Moreover, because of the importance of the labor input in the production process, the quantitative importance of this source of market failure may overshadow the more commonly recognized environmental and other externalities.

## VI. Conclusion

The model of the production process based on the extraction of labor from labor power thus provides an internally consistent micro-economic theory capable of supporting some of the most fundamental general propositions in Marxian economics concerning the reserve army of the unemployed, the determination of the profit rate, discrimination, and the irrationality of the organization of work and technology. The above arguments do not, of course, establish the superiority of the Marxian model. Nor do they provide any indication that the Marxian model is capable of generating plausible empirical accounts of such phenomena as movements in the unemployment rate, the profit rate, the structure of discrimination, or technical choice.

However, a significant amount of empirical work along the lines outlined above has been done, some of it with quite successful results. For example, econometric models of postwar U.S. productivity growth, the profit rate, Tobin's  $Q$ , and strike activity using an empirical measure of the cost of job loss,  $\hat{w}^d$ , have generated highly significant and robust estimates consistent with the expectations of this model.<sup>33</sup> Historical studies of technical choice and work organization based on the extraction of labor from labor

power have produced compelling accounts of otherwise anomalous patterns of technical change. (See Lazonick, 1982, and Marglin.) And econometric studies of the distributional impact of discrimination have produced results quite consistent with the divide and rule interpretation. (See Reich.) None of these is alone decisive, but taken together they do suggest that the Marxian model offers a promising direction for empirical investigation.

## REFERENCES

- Akerlof, George, "A Theory of Social Custom, of Which Unemployment May Be One Consequence," *Quarterly Journal of Economics*, June 1980, 94, 749-75.
- Alchian, Armen and Demsetz, Harold, "Production, Information Costs and Economic Organization," *American Economic Review*, December 1972, 52, 777-95.
- Becker, Gary, "Investment in Human Capital," *Journal of Political Economy*, October 1962, 70, 9-49.
- Boddy, Raford and Crotty, James, "Class Conflict and Macro Policy: The Political Business Cycle," *Review of Radical Political Economics*, Spring 1975, 7, 1-19.
- Bowles, Samuel, "Competitive Wage Determination and Involuntary Unemployment," mimeo., 1981.
- \_\_\_\_\_, (1983a) "Long Term Growth and Equilibrium Unemployment in an Open Competitive Capitalist Economy," mimeo., 1983.
- \_\_\_\_\_, (1983b) "The Cyclical Movement of Real Wages, Labor Productivity, and 'Overhead Labor' in a Competitive Non-Monetary Economy," mimeo., 1983.
- \_\_\_\_\_, and Gintis, Herbert, *Schooling in Capitalist America*, New York: Basic Books, 1976.
- \_\_\_\_\_, Gordon, David M. and Weisskopf, Thomas, "The Profit Rate in the Post-War U.S. Economy: An Econometric Investigation," mimeo., 1983.
- Braverman, Harry, *Labor and Monopoly Capital*, New York: Monthly Review Press, 1974.
- Calvo, Guillermo, "Quasi-Walrasian Theories

<sup>33</sup>See Thomas Weisskopf et al. (1983), my paper with Gordon and Weisskopf (1983), and Schor's and my papers (1983, 1984). Michele Naples (1982) has estimated significant relationships between labor productivity and the structure of control of the labor process consistent with the above model.

- of Unemployment," *American Economic Review Proceedings*, May 1979, 69, 102-07.
- Coase, Ronald, "The Nature of the Firm," *Economica*, November 1937, 4, 387-405.
- Commons, J. R., *History of Labor in the United States*, New York: Macmillan, Vols. 1, 2, 1918; Vols. 3, 4, 1935.
- Eaton, B. Curtis and White, William D., "Agent Compensation and the Limits of Bonding," *Economic Inquiry*, July 1982, 20, 330-43.
- Edwards, Richard C., *Contested Terrain: The Transformation of the Workplace in the Twentieth Century*, New York: Basic Books, 1979.
- \_\_\_\_\_, Gordon, David M. and Reich, Michael, *Segmented Work, Divided Workers: The Historical Transformation of Labor in the United States*, Cambridge: Cambridge University Press 1982.
- Gintis, Herbert, "The Nature of the Labor Exchange," *Review of Radical Political Economics*, Summer 1976, 8, 36-54.
- \_\_\_\_\_, and Ishikawa, Tsuneo, "Wages, Work Discipline, and Macroeconomic Equilibrium," mimeo., 1983.
- Glyn, Andrew and Sutcliffe, Robert, *British Capitalism: Workers, and the Profit Squeeze*, London: Penguin, 1972.
- Goodwin, Richard, "A Growth Cycle," in C. H. Feinstein, ed., *Capitalism and Economic Growth*, Cambridge: Cambridge University Press, 1967.
- Kalecki, Michel, "Political Aspects of Full Employment," *Political Quarterly*, October-December 1943, 14, 322-30.
- Katzell et al., R. A., *Work, Productivity and Job Satisfaction: An Evaluation of Policy Related Research*, New York: Harcourt Brace Jovanovich, 1975.
- Kohn, Melvin, *Class and Conformity, A Study in Values*, Homewood: Dorsey Press, 1969.
- Lange, Oskar, "Marxian Economics and Modern Economic Theory," *Review of Economic Studies*, June 1935, 2, 189-201.
- Lazear, Edward, "Agency, Earnings Profiles, Productivity, and Hours Restrictions," *American Economic Review*, September 1981, 71, 606-20.
- Lazonick, William, "The Subjugation of Labor to Capital: The Rise of the Capitalist System," *Review of Radical Political Economics*, Spring 1978, 10, 1-31.
- \_\_\_\_\_, "Technological Change and the Control of Work: A Perspective on the Development of Capital Labor Relations in U.S. Mass Production Industries," Discussion Paper No. 821, Harvard Institute of Economic Research, 1981.
- \_\_\_\_\_, "Production, Productivity and Development, Theoretical Implications of Some Historical Research," Discussion Paper No. 876, Harvard Institute of Economic Research, 1982.
- Malcomson, James, "Unemployment and the Efficiency Wage Hypothesis," *Economic Journal*, December 1981, 91, 848-66.
- Marglin, Stephen, "What Do Bosses Do? The Origins and Function of Hierarchy in Capitalist Production," *Review of Radical Political Economy*, Spring 1974, 6, 60-112.
- Marx, Karl, *Capital*, Vol. I, Harmondsworth: Penguin, 1976.
- \_\_\_\_\_, *Capital*, Vol. III, New York: International Publishers, 1967.
- \_\_\_\_\_, *The Grundrisse: Introduction to Critique of Political Economy*, New York: Vintage, 1973.
- Medoff, James and Abraham, Katherine, "Experience, Performance, and Earnings," *Quarterly Journal of Economics*, December 1980, 95, 703-36.
- Miyazaki, Hajime, "Work Norm and Involuntary Unemployment," Discussion Paper No. 7, Stanford Workshop on Factor Markets, July 1981.
- Naples, Michele, "The Structure of Industrial Relations, Labor Militance and the Rate of Growth of Productivity: The Case of U.S. Mining and Manufacturing, 1953-1977," unpublished doctoral dissertation, University of Massachusetts, 1982.
- Negishi, Tekashi, *Microeconomic Foundations of Keynesian Macroeconomics*, Amsterdam: North-Holland, 1979.
- Okun, Arthur, *Prices and Quantities, A Macroeconomic Analysis*. Washington: The Brookings Institution, 1981.
- Reich, Michael, *Racial Inequality*, Princeton: Princeton University Press, 1980.
- Roemer, John, "Divide and Conquer: Micro Foundations of the Marxian Theory of Discrimination," *Bell Journal of Economics*, Autumn 1979, 10, 695-705.
- Samuelson, Paul, "Wage and Interest: A Mod-

- ern Dissection of Marxian Economic Models," *American Economic Review*, December 1957, 47, 884-912.
- Schor, Juliet and Bowles, Samuel, "Conflict in the Employment Relation and the Cost of Job Loss," mimeo., 1983.
- \_\_\_\_\_ and \_\_\_\_\_, "Employment Rents and Class Conflict: An Empirical Investigation," mimeo., 1984.
- Shapiro, Carl and Stiglitz, Joseph, "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review*, June 1984, 74, 433-44.
- Solow, Robert, "Alternative Approaches to Macroeconomic Theory: A Partial View," *Canadian Journal of Economics*, August 1979, 12, 339-54.
- \_\_\_\_\_, "On Theories of Unemployment," *American Economic Review*, March 1980, 70, 1-10.
- Weisskopf, Thomas, Gordon, David M. and Bowles, Samuel, "Hearts and Minds: A Social Model of U.S. Productivity Growth," *Brookings Papers on Economic Activity*, 2:1983, 381-450.
- Williamson, Oliver, "The Organization of Work," *Journal of Economic Behavior and Organization*, March 1980, 1, 5-38.



# The Validity of Profits-Structure Studies with Particular Reference to the FTC's Line of Business Data

By GEORGE J. BENSTON\*

The simple proposition that consumer-damaging collusion is more likely to occur when there are fewer competitors has given rise not only to legal restrictions of economic activity thought to restrict output, but also to an enormous amount of empirical work attempting to relate market concentration to the exercise of monopoly power.<sup>1</sup> Most such studies measured market structure by an index of concentration (for example, four-firm concentration or Herfindahl) and performance by accounting profit (for example, net profit divided by assets) or price-cost margins (sales less direct costs divided by sales). From the positive, statistically significant correlations often found between greater concentration and profit (with other factors presumably accounted for), some researchers conclude that increases in concentration are anticompetitive and that concentration, therefore, is bad.

The concentration-profits studies have been criticized essentially on two grounds. One is that the positive relationship is not indicative of collusive behavior. In particular, Oliver Williamson (1968), John McGee (1971), Harold Demsetz (1973, 1974), and Sam Peltzman (1977) point out that higher profits could result from efficiencies experi-

enced by large firms, which resulted both in greater market shares and high levels of concentration. Analytical arguments show that higher profits logically follow as much from lower costs that are associated with higher levels of concentration, as from collusion-determined higher prices. Some empirical evidence that supports this belief is presented by Demsetz, Peltzman, and Bradley Gale and Ben Branch (1982). However, their findings have been contested on the grounds that the data used are biased or inadequate, or that the researchers have not demonstrated that the observed higher-profits/greater-concentration relationship was caused by lower costs.<sup>2</sup>

The second criticism is that available data provide an inadequate basis for such conclusions. The concentration numbers are based, usually, on industries defined by the Commerce Department's Standard Industrial Classifications (SIC). The SIC definitions tend to be supply (production) rather than demand determined, include nonhomogeneous products, and exclude sales of similar products that are included in different SIC groups or are imported. The profit data are taken from accounting reports that provide poor measures of economic values. Weiss (1974) describes many of these problems. However, he does not believe that these shortcomings invalidate the studies. As Weiss (1979) concludes:

I argued that the crudeness of the concentration data, the increasing diversification of firms, and many distortions

\*Professor of Accounting, Economics, and Finance, Graduate School of Management, University of Rochester, Rochester, NY 14627. I am indebted to Armen Alchian, Stanley Liebowitz, and the members of economics workshops at the University of Western Ontario and the University of Rochester Graduate School of Management for valuable suggestions.

<sup>1</sup>See Leonard Weiss (1974) for a review of 46 papers, a presentation of a 47th, and a footnote mention of 8 more, which he notes excludes studies on banking markets. For a review of 15 additional studies on banking markets, see my article (1973). Weiss (1970) also reviews many of these studies but does not find them largely inadequate, as I do. Since the time of these surveys, many more similar papers have been published, of which a few are reviewed below.

<sup>2</sup>See F. M. Scherer (1979b), who shows how the industry output and price statistics used by Peltzman are likely to be biased. Scherer argues that product innovations rather than production efficiencies are causally related to higher profit rates. He does not provide similar (or any) evidence showing that the higher profits are due to collusion.

in accounting profits all bias the observed relationship between concentration and profits towards zero. Because of these biases, I argued that the effect [of a significant positive relationship] was probably understated when observed, and that it might well have been present when it was not detected. [pp. 1106-07]

Largely as a consequence of these problems (particularly the increasing diversification of large companies that made their reported profits and other data difficult, if not impossible, to link with measures of market concentration), the Federal Trade Commission (FTC) designed and implemented a large scale program to gather detailed data from large companies by lines of business (*LB*).<sup>3</sup>

Originally, the program sought to collect data from some 2000 companies on their revenues, expenses, and assets, allocated to 357 lines of business. After vigorous protests, culminating in an unsuccessful 1975 challenge in the Washington, D.C. District Court, the program was scaled back somewhat, and 450 of the largest industrial companies were required to file reports on 267 FTC-designated lines of business (that follow the SICs) for 1974 through 1977. (Partial reports also were filed for 1973.) These data are much more extensive and detailed than have ever been collected by a public agency. Through September 1983, the FTC published five *Statistical Reports*, and forty papers and reports based on the data have been written by its staff and outside economists.<sup>4</sup> Thus, there is a substantial basis on which the usefulness of the program, and of profit-structure studies in general, can be judged. Considering the large number of observations, the detail, the cost of the data, and the qualifications of the

economists who used them, an analysis of the inputs and outputs of the *LB* program can serve as a general critique of the very large body of empirical work that is based on company accounting numbers and data classified by SICs.<sup>5</sup> It also can put into context the large number of studies and other analyses that have been and are likely to be generated from the FTC's *LB* data, since these numbers are on computer tape and are being made available to approved researchers and published in tabular form by the FTC.

### 1. The *LB* Program's Purposes

The first *Statistical Report* published under the FTC's *LB* program reiterates the purposes for which *LB* data were to be used.<sup>6</sup> According to the FTC staff, "The Line of Business Program is designed to elicit information vitally needed in evaluating industry performance" (FTC, 1981a, p. 2). These data, they say, will serve the following purposes:

1. *Antitrust Enforcement*: "Knowledge of industry performance is essential to efforts of the Commission to direct its activities towards industries where poor performance suggests the need for more detailed investigation and possible enforcement action."

[p. 3]

2. *Allocation of Private Sector Resources*: (a) "Managers and directors of corporations will be able to evaluate the performance of their own enterprises against industry averages;" (b) "investment analysts and investors are likely to find *LB* reports useful in evaluating the prospects of particular industries."

[p. 4]

3. *Aid for Economic Research in Industrial Organization*: "And, like economists within the FTC, outside scholars

<sup>3</sup>A resolution of the alternative (efficiency) explanation of the data might also have been a goal, but it was not mentioned in the FTC's Statements of Purpose. See my article (1979, pp. 61-65) for a review of these statements. Also see Federal Trade Commission (1981a, pp. 1-7), which repeats the purposes of the program and does not explicitly mention the efficiency-profits-structure hypothesis.

<sup>4</sup>At least 19 of these have been published or accepted for publication.

<sup>5</sup>From 1978 through 1981, 112 econometric studies in industrial organization were published, according to a review of the *Journal of Economic Literature* and the Index of the *Harvard Business Review* (William Long et al, 1982, p. 141; the papers are listed in their appendix D).

<sup>6</sup>See FTC, 1981a, pp. 1-6.

will use the Line of Business aggregates as a basic data source for advancing the frontiers of knowledge in the field of industrial organization." [pp. 4-5]<sup>7</sup>

It is clear that the *LB* program requires numbers that adequately reflect the desired *economic* market values. Moreover, to be useful for the purposes delineated, the economic market values would have to separate the present values of gains from monopoly power from those due to other causes, such as luck or disequilibrium.

## II. The Usefulness of the FTC's Line of Business and Similar Data for Evaluating Economic Performance

The validity of company accounting data for economic analyses, such as those to which the FTC's *LB* program is directed, would be best determined with tests of prediction. However, these require a measure of "true" economic value, which (to my knowledge) is not available. Hence, two procedures (followed in Sections II and III) are employed. In Section II, accounting numbers, as they are reported to users of published financial statements and to the FTC, are analyzed a priori to determine the extent to which these numbers appear to reflect economic market values. This exercise is undertaken because, as is discussed in Part A, accounting biases are unlikely to be randomly distributed with respect to most of the variables of interest for testing structure-performance hypotheses. The analysis of accounting biases that follows in Part B shows why researchers would find it very difficult, if not impossible, to measure and adjust for the biases, even

were much firm-specific information available. The empirical procedures employed by FTC economists to overcome two problems, overhead allocation and transfer pricing, are considered here. In Part C, the degree to which the data are contaminated and incomplete, given the FTC's (SIC) definition of industries, is assessed.<sup>8</sup> Some evidence supporting the concerns about the data is presented in Part D. In Section III, the application to hypothesis testing of the accounting numbers collected by the FTC is evaluated, to determine whether the researchers were able to overcome the severe shortcomings of the data that are delineated in Section II.

### A. The Interaction of Accounting Measurement Biases and Measures of Performance

The analysis presented below shows that differences between accounting measures and economic market values are likely to be significant and very difficult (in many important instances, impossible) to determine. However, were they randomly distributed with respect to variables of interest, the result would only be noise or possibly spurious correlations. It is doubtful, though, that such randomness occurs, for the three reasons outlined next.

First, several studies have shown that firms' choices of accounting procedures are functions of their size, propensity to be regulated, management profit-sharing plans, debt covenants, and tax situation. For example, Ross Watts and Jerold Zimmerman (1978) find the larger firms and those more likely to be regulated tend to choose accounting methods that result in lower reported profits. Robert Hagerman and Mark Zmijewski (1979) find that systematic differences in the accounting

<sup>7</sup>In the FTC's previous releases, which also stress these purposes, mentioned additionally are improvements in the government's efforts to control inflation and unemployment, aids to small business firms and to buyers and sellers of goods, facilitation of labor unions' bargaining by providing them with information on profits, and information to farm groups that will help them in dealing with suppliers and processors. For references to the relevant documents and critique of the aims of the program, see my article (1979). Also see William Breit and Kenneth Elzinga (1981). For an opposing view, see Scherer (1979a).

<sup>8</sup>The poor conformance of four-digit SIC-defined markets (which the FTC uses to define lines of business) to homogeneous industries is discussed and demonstrated by Weiss (1974, pp. 194-96), Betty Bock (1975), and my article (1979, pp. 66-78), among others. (See fn. 49 below for a very brief description of five such "industries.") Therefore, though this problem is severe, it is not analyzed further here. An additional bias is the restriction of the program to large firms.

methods adopted are related to whether the firms are management controlled or owner controlled. Robert Holthausen and Richard Leftwich (1983) report that the presence of debt covenants affected the accounting procedures used. Unfortunately for researchers who would use accounting data, these biases need not be constant. For example, if a firm is in a steady state, profit understated in an earlier year results in an overstatement in a later year, *ceteris paribus*. Other combinations of effects can occur, as is discussed below. Thus, even if these biases were known and identified with specific firms, the researcher must also deal with the effects of past decisions to adopt alternative accounting procedures.

Second, alternative accounting procedures need not be consistently applied by all firms in an industry. This is likely to be a greater problem when industries are defined by SIC codes. It is exacerbated further when industries are formed by aggregating the FTC-defined *LBs* of different firms, firms with accounting practices determined by their original industry status. For example, Eastman Kodak may have adopted firm-wide accounting methods that are appropriate for a film and camera maker, but its chemical activities are aggregated by the FTC with those of, say, Union Carbide which may use methods appropriate for a chemical manufacturer that specialized in batteries. Additionally, centralized firms that do not allocate expenses, assets, and capital to product lines may be compared with decentralized firms. As a consequence the former report higher price-cost margins ( $P/S$ ) and returns on assets ( $P/A$ ), *ceteris paribus*, where costs are defined as allocated expenses. Furthermore, to the extent that accounting biases are industry related, observations from different industries cannot be compared or pooled without the differences being adjusted for.

Complicating matters further, the FTC's rules permit firms leeway in assigning their data to specific SIC codes. Consequently, accounting differences among a changing sample could be responsible for the considerable changes (reported in Part D below) in the aggregate industry ratios published by

the FTC in its *Statistical Reports* (1981a,b; 1982). These are the data that will be available to non-FTC researchers and other users. The potential for drawing incorrect inferences from these data would appear to be considerable.

Third, there is reason to believe that accounting mismeasurement of economic market values is associated with some commonly used measures of company performance. For example, a concentrated industry may be one in which only a relatively few firms compete because a large investment in fixed assets is required. If the values of these assets are understated, perhaps because they are recorded at amounts that do not reflect inflation, depreciation also is understated. Consequently, such measures as  $P/S$  and  $P/A$  will be positively associated with concentration, *ceteris paribus*. Similarly, if a few companies dominate an industry because they previously invested in intangibles (such as patents, industrial processes, marketing networks, and advertising), the current reported  $P/S$  and  $P/A$  will be overstated. Thus market share and profit rates will be positively associated, *ceteris paribus*, though not as a consequence of market power. Indeed, if the amounts expended on intangibles were correctly accounted for, it might be that the acquisition of a large market share was a negative present value investment. Unfortunately, this bias need not be stable. For example, the reported  $P/S$  of a growing company could be understated. (These possibilities are discussed further in subsection B1.)

As a further example, it may be that executives who manage lines of business with large market shares are compensated, in part, with a share of accounting profits. In a particular year, they (and their bosses) may find it desirable to show larger profits. Or companies with *LBs* that show large market shares may be those that experienced a surge in demand, perhaps because the prices of substitutes increased. If these companies allocated overhead to product lines as a fraction of direct costs, their product costs would not increase as much as their revenues. Hence  $P/S$  and  $P/A$  would be positively associated with higher market shares, *ceteris paribus*.

But the relationship would not be as positive had they allocated overhead according to sales dollars. In either event, a measured association between profits and market shares would reveal nothing about market power or efficiency.

Many more examples could be given to explain measured relationships between accounting profits and concentration or market share. Therefore, researchers who would use company accounting data should at least attempt to adjust for biases or show that these are not important. On the first point, the following analysis demonstrates that the required adjustments are very difficult or (in many important situations) impossible to effect. Though some researchers who use accounting numbers profess awareness of the inherent biases, they do not seem to realize just how complex the situation is.

#### B. *Company Accounting Data and Economic Market Values*

The extent to which accounting data differ from economic market values is presented in subsection 1. The usefulness to companies of these data, despite their considerable divergencies from economic market values, is explained in subsection 2.

1. *Divergencies of Accounting Data from Economic Market Values.* For present purposes, only the generally most important sources of divergencies are considered.<sup>9</sup> These include: (a) the values of long-lived tangible assets and their depreciation or depletion over time; (b) the value and amortization of intangible assets (for example, advertising and research and development); (c) inventory values and the cost of goods sold; (d) allocation of common and joint costs to lines of business; and (e) intrafirm transfers among lines of business. For each, the reasons for the divergencies are delineated first, followed by an evaluation of the extent to which the reported numbers can be adjusted to approximate economic market values and the

effect of the remaining biases on two commonly used measures of economic performance, the profit rates on sales and assets. The conclusions of the analysis are summarized in Table 1 and an analytic estimate of the cumulative impact of the accounting biases is given in subsection (f).

#### (a) *Long-Lived Tangible Assets and their Depreciation or Depletion Over Time.*<sup>10</sup>

*Accounting Procedures.* Accountants do not attempt to record, initially, an asset's economic value to a firm, but, rather, the amount paid for it as measured in objectively determined monetary units (market value). An asset's economic or present value (value in use) to the firm should be greater than or equal to its market value (value in exchange), since otherwise the asset would not have been purchased. The difference could be due to nonacceptable (for example, illegal monopoly power) or acceptable (for example, specialized use of resources) rents. Neither is recorded at the time of the transaction.

Where an asset is constructed rather than purchased, the recorded amount is likely to understate its competitive market value, since the seller's selling and administrative costs and total cost of capital rarely are recorded. But an asset is never capitalized (recorded as an investment) at more than its recorded cost, and, when the asset could have been purchased at less than the amount expended to make it, accountants charge the excess to expense (the conservative accounting bias).

Understatements of initial asset values are particularly serious for extractive assets. When these assets are fortuitously discovered, at most only the amounts expended for the successful discovery are capitalized. In other cases, the successful efforts method of accounting (used by most larger natural resource producers) capitalizes only the direct cost of discovering and developing successful wells and mines. The amounts spent for unsuccessful efforts are written off (expensed) as current expenses, despite the fact that successful ventures usually are accompa-

<sup>9</sup>For a much more detailed and differently structured and directed analysis, see my article (1982).

<sup>10</sup>The discussion in this subsection applies generally to the following subsections. Hence it is relatively long.

TABLE 1—THE EFFECT OF SOME DIVERGENCIES BETWEEN ECONOMIC VALUES AND ACCOUNTING NUMBERS ON THE RATIOS OF PROFIT TO SALES ( $P/S$ )<sup>a</sup> AND PROFIT TO TOTAL ASSETS ( $P/A$ )<sup>a</sup>

	Initial Period		Later Periods	
	$P/S$	$P/A$	$P/S$	$P/A$
Capital Gains—One Time				
Initially recorded assets and increases in present values	<i>UU</i>	<i>U</i>	<i>O</i>	<i>OO</i>
Capital Gains and Overdepreciation—Multiple Events				
Stable (uniform distribution of events)			<i>N</i>	<i>O</i>
Declining			<i>O</i>	<i>OO</i>
Growing			<i>U</i>	<i>?</i>
Expected Inflation	<i>N</i>	<i>N</i>	<i>O</i>	<i>OO</i>
Unexpected Inflation	<i>?</i>	<i>?</i>	<i>O</i>	<i>OO</i>
Intangible Assets				
Stable (current expenditures equal depreciation)			<i>-</i>	<i>O</i>
Declining			<i>O</i>	<i>OO</i>
Growing			<i>U</i>	<i>?</i>
Cost of Goods Sold and Inventories				
Prices increasing steadily:				
Last-in, First-out			<i>N</i>	<i>O</i>
First-in, First-out			<i>O</i>	<i>O</i>
Prices decreasing steadily:				
Last-in, First-out			<i>N</i>	<i>U</i>
First-in, First-out			<i>U</i>	<i>U</i>
Overhead allocations among products			<i>?</i>	<i>?</i>
Allocation of Joint and Common Costs and Assets to <i>LBs</i>			<i>?</i>	<i>?</i>
Intrafirm Transfers Among <i>LBs</i>				
Contamination due to Assignments to <i>LBs</i>			<i>?<sup>b</sup></i>	<i>?<sup>b</sup></i>
Transfers priced at other than market			<i>?<sup>b</sup></i>	<i>?<sup>b</sup></i>

Notes: *U* = understated; *UU* = very understated; *O* = overstated; *OO* = very overstated; *N* = not affected; *?* = bias unknown.

<sup>a</sup>Profit is before interest and income taxes.

<sup>b</sup>Probably not very great on average.

nied by unsuccessful ones. Even companies that use the full-cost method (under which amounts incurred for all discovery efforts are capitalized) expense many development costs.

Once the tangible assets have been acquired, the amounts recorded as asset values are expensed over the expected economic life of the assets. The tax regulations usually permit a faster write off, which may also be used for book (reporting) purposes. But a book method may be adopted that yields currently higher reported net income. In any event, the depreciation method chosen is not designed to measure the user cost of the assets—the change in the asset's present values—in part because these calculations are too difficult to make, and in part because the

accountants who must attest to the numbers would find such calculations uncomfortably subject to manipulation.

The recorded numbers also are not adjusted to reflect changes in the capital values of the tangible assets. Such changes could occur as shifts take place in the supply and demand for the assets, the factors used with the assets for production, and the goods produced.<sup>11</sup> When the changes are due to monopoly power or collusive pricing, the

<sup>11</sup>For example, the 1970's increase in the market price of oil increased the values of oil and gas reserves, substitutes such as coal, and fixed-price contracts to purchase these resources.

economist interested in these events would not want the values recorded as ordinary assets, since all firms then would be measured as earning zero profits. Therefore, the reasons for the change in values would have to be specified and their effects measured.<sup>12</sup> In any event, the accounting values are not changed.

The monetary values of assets also change as the purchasing power of money changes. While estimates of the effect of general price level changes now appear in the published statements of large companies, usually by the application of indices, these adjustments are not often made a part of a company's accounting system and individual asset accounting numbers are rarely changed. Furthermore, current values are not recorded, largely because the cost of such adjustments is greater than their value to most companies.

The magnitudes of the divergences between accounting numbers and economic market values cannot be determined.<sup>13</sup> An estimate, perhaps, could be made if one could obtain estimates of the current market values of individual assets. Even then, the assets' economic (present) values to the firm would not be measured, nor would the extent to which the present values are due to presumably unacceptable (for example, collusive pricing) or acceptable (for example, entrepreneurial skill) behavior. Furthermore, market-revealed values are unlikely to occur for most user-constructed assets. Even general adjustments in the magnitudes, perhaps with some index, would require detailed information about purchase dates, current

equivalents in terms of production, and the economically correct depreciation procedure to be applied. Such knowledge is impossible or very expensive to obtain.

*The Effect of the Biases on Measures of Economic Performance.* Although there is no practical way to convert accounting asset values and depreciation to economic values or to separate the causes of changes in values, those who want to use these numbers claim they could be useful if the direction and rough magnitude of the biases could be stated. But, as the following analysis shows, the effect of the biases on two commonly used ratios—operating profit to sales ( $P/S$ ) and operating profit to total assets ( $P/A$ )—is not determinable.<sup>14</sup>

As noted, increases in the value of long-lived assets are not recorded, resulting in understated profit and asset values. Consequently  $P/S$  and  $P/A$  are understated (assuming  $P/A$  is less than 1.0). Subsequently, profit and  $P/S$  and  $P/A$  are overstated, since depreciation on the higher unrecorded asset value is not recorded, and the asset remains understated. Unrecorded decreases in assets values have the opposite effect.

The effects become more complicated when the initial events and changes in asset values occur more than once, as is likely. Even if the events occurred uniformly over time, and

<sup>12</sup>The effect of monopoly on asset values could be (and has been) measured by relating changes in the market value of assets to changes in laws or firm behavior (for example, mergers) that are hypothesized to result in monopoly. These studies include taxicab licensing, stock exchange regulation, and lawsuits filed by the Department of Justice. None of these studies use accounting data. See G. William Schwert (1981) for references and a discussion of the methodology.

<sup>13</sup>Some economists (such as James Buchanan, 1969, and others whose essays are reprinted in Buchanan and G. F. Thirlby, 1973) believe that economic values (particularly costs) cannot be measured conceptually, since they depend on subjective evaluations of alternatives.

<sup>14</sup>See Ezra Solomon (1970) and Thomas Stauffer (1971) for formal proofs. Franklin Fisher and John McGowan (1983) show that the divergence between accounting and economic rates of return are likely to be nonpredictable and nontrivial. Gerald Salamon (1973) criticizes researchers who attempted to relate the firm accounting rate of return ( $P/A$ ) and internal rate of return ( $IRR$ ) by assuming that the firm  $IRR$  is equal to the  $IRR$ s of the firm's projects. Salamon shows that "...if the growth rate in gross investment ( $g'$ ) is greater than or equal to the  $IRR$  of firm projects ( $r'$ ) then the  $IRR$  of the firm is not equal to the  $IRR$  of firm projects; and in fact, the  $IRR$  of the firm is not even defined" (p. 301). Salamon (1985) also shows that an approximation to firm  $IRR$  derived from net cash flows is not significantly related to firm size, while  $P/A$  is significantly positively related. The measure he employs, though, is based inter alia on an assumed constant growth rate in gross investment and on exponentially decaying (or growing) project cash flows. He warns: "Whether the measurement error from this source [the assumed cash flows] is large or small, systematic or random is pure conjecture at this point."

the effects on profit washed out, assets would continue to be understated, and  $P/A$  would be overstated (assuming unrecorded capital gains). But if the events diminished over time, perhaps because of a decline in growth of the company or in relative prices, unrecorded depreciation probably would exceed unrecorded capital gains. Hence profit would be overstated. The cumulative understatement of assets (which would not be removed until disposal of the assets) would result in a greater overstatement of  $P/A$  than in  $P/S$ . But if the company were growing or relative prices were increasing, unrecorded capital gains probably would exceed the unrecorded depreciation of past unrecorded capital gains. Thus  $P/S$  would be understated. The effect on  $P/A$  is uncertain, since both the numerator and the denominator of the ratio are understated.

The effect of general price level changes is different because expected changes do not give rise to economic income or expense. But if the recorded nonmonetary asset amounts are not changed to reflect the current purchasing power of the dollar, an expected inflation results in understated assets and overstated profit (since depreciation is understated). Therefore, with respect to nonmonetary assets,  $P/S$  is overstated and  $P/A$  even more so. Interest income and expense also are overstated since they include a return on capital. (If the interest expense overstatement were large enough, which is doubtful,  $P/S$  could be understated.) Where the inflation is unexpected, though, the company realizes an unrecorded capital loss or gain if it holds nominally denominated assets or liabilities. Thereafter, its interest income, interest expense, depreciation, and assets are understated. The effects on  $P/S$  and  $P/A$  thus depend on the extent to which it holds nominally denominated assets and liabilities. If the unrecorded capital gains exceed the capital losses, the first-period  $P/S$  is understated;  $P/A$  is understated if the unrecorded gains were large enough. In subsequent periods,  $P/S$  and  $P/A$  are overstated (unless the unrecorded interest income is large enough).

While these biases are generally as specified, the magnitude of the effect varies with several factors. Profit rates and rates of re-

turn of companies with longer-lived assets are more affected than those with shorter-lived assets. The amounts invested in long-lived compared to other assets similarly affects the bias. Profit rates are overstated more for rapidly growing than for slowly growing companies. The methods used by the company to depreciate or deplete its assets (for example, straight-line, sum-of-the-years' digits, double declining balance, or units of production) also affects the bias. The times at which assets yield cash flows and the relevant discount rate also affect the difference between the economic profit and rate of return and the accounting numbers. Furthermore, the extent to which the initial period effects occur discontinuously affects the bias.

As is explained above (and summarized in Table 1), except in growth situations the bias in the initial period is the reverse of the bias in subsequent periods. Hence, a large enough change in one period can reverse the sign of the continuing effects of previous periods. In general, then, there is no way to know the sign of the difference between the accounting numbers and economic market values. But the differences are likely to be quite large unless the amount of long-lived tangible assets is relatively small, even without inflation. With inflation, and the nonadjustment of accounting numbers (which especially affect long-lived assets purchased at different price levels), the biases are likely to be considerable. Indeed, a sufficiently high inflation would swamp the other sources of divergence between accounting numbers and economic market values. At the least this means that companies with more tangible assets and assets that are longer lived are more likely to overstate profits and rates of return on assets. In the absence of this overwhelming inflationary effect, the direction of the bias cannot be stated a priori, nor does there appear to be any practical way to adjust the reported  $LB$  and other company accounting data to reflect desired economic market values.

#### (b) *The Value and Amortization of Intangible Assets.*

*Accounting Procedures.* The amounts spent for intangible assets—such as advertising,



goodwill, research and development, and employee training—are recorded as current-period expenses in accordance with the “generally accepted accounting principles” of the Financial Accounting Standards Board. Although a large portion of these expenditures clearly results in assets, the amounts are not recorded as such because the present values of the resulting assets are difficult to measure objectively, and managers could use the estimates to manipulate reported profits and assets. Even when estimates were made objectively, the likelihood that they would turn out to be wrong is considerable. Subsequently, even honest estimates could appear to have been deliberately misstated to fool users of the company’s financial statements. To protect the independent auditors who attest to the statements, and as a means of demonstrating that auditable measures of the present values of intangible assets cannot be made, all expenditures on such assets (i.e., intangibles) are recorded as expenses by accounting fiat.

*The Effect of the Biases on Measurements of Economic Performance.* The effect on recorded profit of not capitalizing intangibles depends on their past and current amounts and the rate at which they depreciate. Several economists have attempted to estimate the numbers, an exercise that requires some heroic assumptions. For example, Weiss (1969) assumed a constant annual growth rate of advertising and a constant rate at which past advertising expenditures depreciate. Robert Ayanian (1975), using those assumptions, showed that the rate at which advertising is assumed to depreciate is critical for calculating the economically correct rate of return. He estimated the depreciation rate by assuming a multivariable Cobb-Douglas production function for advertising. To estimate the parameters, ten years’ data were required. William Comanor and Thomas Wilson (1979) dismiss his results because they believe that advertising generally depreciates very rapidly. But the studies they cite are no better than Ayanian’s in measuring the depreciation rate because a basic characteristic of intangibles such as advertising is the difficulty, if not the impossibility, of estimating the value of expendi-

tures, future as well as current, even when one has complete knowledge about the company, its environment, and the specific projects and products to which the expenditures pertain. The issue is completely intractable when one must work with data aggregated into SIC-defined “industries.”

Given the caveats noted above, several effects of the divergence between accounting numbers and economic values on the profit rates on sales and assets ( $P/S$  and  $P/A$ ) can be specified. If the company is in a stable, zero growth state, such that the amount expended on intangibles exactly equals the depreciation of past, implicitly capitalized amounts,  $P/S$  is not affected but  $P/A$  is overstated, because past expenditures were not capitalized into assets. If the company’s expenditures on intangibles has been declining steadily and the depreciation rate of the intangibles is constant, profit is overstated because the decline in value of past written-off expenditures is greater than the amounts currently charged to expense; thus  $P/S$  is overstated. Since assets are understated as long as the past expenditures on intangibles have present value, the overstatement of  $P/A$  is even greater. When expenditures on intangibles are increasing and the depreciation rate is constant,  $P/S$  is understated. But, since assets also are understated, the effect on  $P/A$  is not determinable without additional information. Finally, when the growth rate of intangibles equals the real rate of return,  $P/A$  equals the real rate of return.<sup>15</sup>

The amount of intangible assets relative to tangible assets affects the magnitudes of these biases. Importantly, when the assumption of a constant rate of depreciation is false, neither the magnitudes nor the direction of the bias can be determined.

### (c) *Inventory Values and Cost of Goods Sold.*

*Accounting Procedures.* Accountants assign values to inventories according to several rigidly applied formulas that need not de-

<sup>15</sup>Ayanian’s discussion and illustrations show how difficult it is to make any a priori statements, even when constant rates of growth and decline of intangibles are assumed. Solomon also presents an excellent discussion.

pend on the physical movement of the goods. The first-in, first-out method (*FIFO*) assumes that the first goods purchased are the first sold or used in production; last-in, first-out (*LIFO*) makes the opposite assumption. An average of the prices paid is also an acceptable method. Or, when the items in inventory are relatively few and unique, the acquisition prices of the specific goods sold may be expensed as the cost of the goods sold. For all methods, the value of the goods in inventory should not exceed their market values, net of disposal cost. Thus, the amounts considered to be current expenses and the amounts treated as investments in assets (inventory), depends on the accounting method employed and the pattern of prices paid for the goods. None of the acceptable accounting methods records the cost of goods sold at their opportunity costs—the amounts foregone as a consequence of having sold the goods. In most (though not all) situations, this is the replacement cost of the goods. Nor, except possibly in supplementary schedules, are the goods in inventory valued at their market prices.

Several accounting procedures determine the amounts recorded for goods that a company manufactures. In addition to the conventions governing the cost of materials discussed above, some manufacturers use “standard-costing” to assign numbers to manufactured inventories. The standards applied differ among companies. Overhead, especially, is allocated to products according to several alternative criteria. These are usually arbitrary either because there is no conceptually meaningful way to assign costs that do not vary with output or are joint among outputs, or because the cost of determining the association between cost and output exceeds the benefits thereof. Furthermore, to the extent that overhead or direct costs include such accounting numbers as depreciation, the amounts charged to inventories and then to cost of goods sold diverge further from economic market values.

*The Effect of the Biases on Measurements of Economic Performance.* Short of a revaluation to opportunity costs of inventories and cost of goods sold, there is no way that the reported numbers can be adjusted to

economic market values. However, when the prices of materials and other costs of production are stable and equal to opportunity costs, the only important problems are the divergence of internal accounting numbers from economic market values (for example, depreciation) and the arbitrary allocation of overhead to specific goods and processes (unless a company produces only one good or product line in a plant).

When prices are changing, the method of assigning values to the goods sold and to those remaining in inventories becomes important. If prices are increasing steadily, the *LIFO* method will come the closest to recording costs of goods sold at economic market values. But the inventory will be understated. Hence, as long as the physical inventory does not decrease,  $P/S$  may be reasonably unbiased; but  $P/A$  is overstated. The *FIFO* method tends to value assets at current prices while profit is overstated. Thus both  $P/S$  and  $P/A$  are overstated. When prices are decreasing steadily, the reverse occurs—the overstatements are then understatements. However, since U.S. companies are legally required to use the same inventory costing method for tax returns as they do for their financial statements, they are likely to use *LIFO* in periods of steadily rising prices and then switch to *FIFO* if they believe that prices will steadily decline. (Taxpayers are permitted one unchallenged switch; they need not use the same costing method for all inventories.) But when price levels fluctuate, the direction of the bias is not unique. The effect of alternative overhead allocations to products is another indeterminate element. This is particularly troublesome for the FTC’s line of business data, since information on overhead allocations at the plant level is not reported.

(d) *Allocation of Joint and Common Costs and Assets to Lines of Business.*

*Accounting Procedures.* Many companies charge costs incurred above the plant level (company-wide costs) to current expense. Others allocate these costs to divisions or plants (rarely to products), using such bases as sales, direct labor cost, labor plus materials (prime) cost, and sales less variable costs

(contribution margin). When company-wide costs do not vary as a consequence of individual outputs or outputs aggregated by lines of business, any allocation method is arbitrary, and profit by product line cannot be determined. The amounts classified as company-wide rather than as plant-level also vary according to the accounting procedures adopted by individual companies. Such considerations as styles of management control, past accounting practices maintained for the sake of consistency, reduction of state taxes, and the degree of decentralized organization dictate the extent to which company-wide costs and assets are assigned to plants and divisions. In addition, the accounting numbers for company-wide assets and costs (such as depreciation of central facilities, general company advertising, and basic research and development) are unlikely to provide valid measures of the economic market costs attributable to products or product lines (as explained above).

*Effect of the Bias on Measures of Economic Performance.* The effect and direction of the divergence between the reported profit and asset amounts after or before allocations of company-wide costs and assets and the economically valid amounts is not determinable. Most enterprises are not merely mutual funds in which individual companies, each of which produces and sells a line of products specified by the FTC, are joined only by common ownership of a few resources general to all product lines. Rather, products usually are produced and/or sold together because they share joint demands, costs, or both. There is no conceptually valid way (with respect to economics) to allocate these joint revenues and costs to the individual lines of business. And, unless an analysis were made of how the common costs varied with individual products or product lines and could be avoided were production abandoned, there is no way of knowing how closely allocations made via some base, such as sales or direct expenses, approximate the correct amounts. Nor is there any way of knowing what portion of the company-wide costs and assets are joint or common, and how much the accounting numbers diverge from economic market values. Hence the effect of the bias of

allocating or not allocating those costs and assets designated as company-wide cannot be determined.

Nevertheless, economists at the FTC argue that, if reported company-wide costs were relatively small, they could be ignored. The FTC's *1974 Statistical Report* (1981a, pp. 62-70) states that 86 percent of the companies reported nontraceable costs.<sup>16</sup> Of these, 62 percent used a single base for making allocations (most used sales), 28 percent used two bases, and 10 percent used three bases. For the entire sample of company *LBs* (each of the FTC-defined lines of business of each reporting company) where a company has more than one line of business, the ratio of nontraceable to the sum of nontraceable and traceable amounts averages 14 percent for total selling and general and administration expenses and 12 percent for total assets. Nontraceable ratios over 20 percent are reported by 26 percent of the company *LBs* for expenses and 18 percent for assets. By *LB* industry aggregates, the weighted average nontraceable ratios are 15 percent for expenses and 13 percent for assets. Ratios over 20 percent are reported for 31 percent of the industries for expenses and 19 percent for assets. Thus, compared to the likely size of net profits, allocated expenses and assets are not negligible.

William Long (1981a) estimated some effects of allocating company-wide costs. (Alternative allocations of company-wide assets were not considered.) He did so by investigating the effect of alternative allocation methods on a regression of profit/sales on seventeen variables, using company *LB* data. Profit originally was calculated as gross profit (sales plus transfers less cost of sales) less direct and nontraceable expenses as allocated by the reporting companies.<sup>17</sup> When nontraceable selling and administrative expenses were reallocated according to sales or a sales-dominated base, the coefficients were not altered much, a result which is not

<sup>16</sup>The 1975 data are similar; see FTC (1981b, pp. 54-59).

<sup>17</sup>The original regression was run by David Ravenscraft (1982).

surprising since 69 percent of the companies allocating common costs used sales as their allocation base and most of the balance included sales in their bases. However, the effect of allocating these costs randomly (which, in the absence of theory or evidence on how the nontraceable expenses vary, has about as much validity as the other methods), Long reports, "is large and disastrous. The F-ratio plummets. Coefficients change signs. Significance levels are altered" (1981a, p. 21).<sup>18</sup>

Additional insights are obtained from an in-depth study of the financial data of six large companies by Robert Mautz and Fred Skousen (1968).<sup>19</sup> They recorded the net profit of company-determined business segments (three to nine segments for each company, thirty in all) as measured by each company's allocations of noninventoriable common expenses (excluding income taxes). Then they calculated the segment net profit after allocating the common expenses according to four additional, commonly used allocation bases (sales, total assets, number of employees, gross profits, and combinations thereof). The common expenses allocated are relatively less than in the FTC data set, averaging 3 percent of total expenses compared to 14 percent for the *LB* data,<sup>20</sup> probably because the company-determined segments tend to be broader than the SIC-

determined *LB*s. Consequently, Mautz and Skousen's allocations probably understate the magnitudes of the effects on net profit that might be found in the FTC data. Even then, the alternative allocation procedures yield very different rates of return (profit/assets). The differences can be summarized by dividing the range of the alternatively calculated returns by the company-determined return for each of the thirty business segments. (In percentages; a quotient of 0 means no difference among the alternatives.) The distribution of these percentage quotients is as follows:

0 to 5 percent—3 segments (10 percent);

6 to 10 percent—4 segments (13 percent);

11 to 20 percent—9 segments (30 percent);

21 to 40 percent—4 segments (13 percent);

41 to 60 percent—2 segments (7 percent);

61 to 80 percent—1 segment (3 percent);

141 to 843 percent—7 segments (23 percent).

With five observations deleted, four from the company with a very low percentage of common expenses and another with a very low company-determined profit rate, there are no observations in the 0–5 percent category, three instead of four in the 6–10 percent category, and six instead of seven in the over 141 percent category.<sup>21</sup>

Thus calculated product-line profit rates appear to be sensitive to the way indirect expenses are allocated, even when commonly used allocation methods are employed. Nor is there any *a priori* reason to favor one allocation method over another or even to know whether any of the methods yield economically meaningful numbers. Rather, the determinants of these expenses within each company should be analyzed empirically, a process that cannot be undertaken with the data requested by the FTC. Even then, where such expenses are jointly determined there is

<sup>18</sup>Long (1981a) also reallocated nontraceable advertising and other selling expenses and recalculated the equations he had calculated for another paper (1980). However, only an average of 3.6 percent of these costs are nontraceable and few (less than 10 percent) of the regression coefficients estimated are significant at the 5 percent level. Hence, though he found few significant changes in the coefficients and other statistics, the finding is not of much value for the present question. Stephen Martin (1981b) uses alternative definitions of total assets (for example, traceable and traceable plus nontraceable) as variables in regressions. Since he does not employ alternative allocation methods for nontraceable items, his results are not relevant here.

<sup>19</sup>Mautz and Skousen state that, although their sample was not randomly selected, "[t]ogether the six companies represent a broad cross-section of American Industry" (p. 19).

<sup>20</sup>One company had common expenses of only 0.5 percent of total expenses. With this company excluded, the average is 3.7 percent

<sup>21</sup>See my article for details (1979, Table 1–2).

no conceptual way to assign them to specific products or product lines.

(e) *Intrafirm Transfers Among Lines of Business.*

*Accounting Procedures.* Intrafirm transfers that are not priced at the opportunity value of the goods impart a mismeasurement of the sales of the sending business unit and the expenses of the receiving unit. For example, when transfers are priced at, say, cost rather than a greater market value, the sending unit's sales and the receiving unit's costs are understated. Because transfers at other than market prices could be repriced only at prohibitive expense, the FTC permitted respondents to report data as per their books, but also asked them to report the transfer pricing method. Transfers from one line of business to another are reported by 80 percent of the companies.<sup>22</sup> Of those transferring among lines of business, the transferred amounts (weighted by sales plus transfers, hereafter denoted as sales) average 8.9 percent of sales. Transfers above 9.9 percent of sales are reported for 24 percent of the company lines of business. Goods transferred to foreign activities (which are not included in the *LB* data) average 3.5 percent of sales (Ravenscraft, 1981, Table I). Thus transfers, while not great on average, are not negligible. Furthermore, differences in tax rates and bases among states and countries affect transfer pricing practices, which impart a systematic bias to the profits data.

The companies reported that they made 39 percent of their transfers at market prices. Cost plus markup was used in 26 percent of the transfers, cost was used for 13 percent, and "other" for 22 percent.<sup>23</sup> About half the

companies used more than one transfer pricing method.

*Effect of the Bias on Measures of Performance.* The size of the bias imparted by intrafirm transfers at other than market prices is analyzed by Ravenscraft (1981). For each company *LB* (3186 observations), he calculated the percentages to sales of transfers out and in that were priced at cost, cost plus markup, and other. These percentages were included as independent variables in regressions of *LB* profitability (profit/sales) on market structure and other variables. The coefficients of these variables show that, where transfers were at other than market, the profitability of the sending *LBs* is significantly lower and the profitability of the receiving *LBs* is significantly higher. Thus, the method of valuing transfers does matter. But he also finds that few of the coefficients of the other variables are affected; for all but two (rather unimportant) variables, the differences are negligible.

Ravenscraft also attempted to correct *LB* profits for transfers at nonmarket prices.<sup>24</sup> This adjustment was only partially successful; several of the measures of the transfers at nonmarket prices remained statistically significant in estimating profit/sales.<sup>25</sup> In comparison with Ravenscraft's original regression, the general conclusions were not altered appreciably, though several coefficients changed significantly.

Ravenscraft reports that the profit/sales measure (which the FTC claims would be useful to analysts) is changed as follows:

The average absolute change in operating income to sales for the 702 *LBs* is .0259, with a maximum change of .2540. In percentage terms, the average change in operating income to sales is 445%, with a maximum change of 17,595%.

<sup>22</sup> FTC (1981a, pp. 46-53). The 1974 data are similar.

<sup>23</sup> Ravenscraft reports sales weighted averages of 50.7 percent (market), 25.2 percent (cost plus markup), 9.4 percent (cost), and 14.7 percent (other). He states that "By far the two most common classifications in 'other' are negotiated value and market, list, or wholesale price less discount" (1981, p. 26, fn. 2). These methods apparently were used in about a third of the "other" lines of business. Though these methods seem to be close to market values, his regressions indicate that, in relation to market price, transfers at "other" prices were lower than transfers at cost or cost plus markup.

<sup>24</sup> The revaluation factor was calculated by dividing those industry sales and transfers that were valued at market by the cost of goods sold plus general and administrative expenses and less the amounts transferred at other than market. By this method, the profits of 22 percent of the *LBs* were reallocated.

<sup>25</sup> Curiously, the coefficients of transfers made at cost and cost plus markup change from significantly negative to significantly positive.

TABLE 2—REPORTED PROFIT RATE ON SALES OR ASSET WHEN COSTS ARE UNDERSTATED OR OVERSTATED BY A PERCENTAGE ERROR<sup>a</sup>

Understatement (–) or Overstatement (+) of Costs	Actual Profit Rate on Sales or Assets				
	0.05	0.10	0.15	0.20	0.30
–0.30	0.34	0.37	0.41	0.44	0.51
–0.20	0.24	0.28	0.32	0.36	0.44
–0.10	0.15	0.19	0.24	0.28	0.39
–0.05	0.10	0.15	0.19	0.24	0.34
+0.05	0.00	0.06	0.10	0.16	0.27
+0.10	–0.05	0.01	0.07	0.12	0.23
+0.20	–0.14	–0.08	–0.04	0.04	0.16
+0.30	–0.24	–0.20	–0.11	–0.04	0.09

<sup>a</sup> Calculated with formula derived in fn. 18.

[This is an average of the fifteen largest changes—so the actual maximum is higher.] The correlations between the original and reallocated profits is .8936. [1981, p. 10]

On the industry level, the two measures of profitability are highly correlated at .99. The average absolute change in the ratios is .0033; the average absolute percentage change expressed as percentages of the ratios is 11.5. But among the 237 industries reported by the FTC, there are some dramatic changes. The 10 industries with the largest changes show an average absolute percentage change of 170 percent, with a maximum of 590 percent. Thus individual profitability numbers can be considerably affected by transfers at other than market prices.

#### (f) *The Cumulative Impact of Accounting Biases.*

The extent to which accounting numbers diverge from economic market values cannot be determined, either as to magnitude or sign, but the effect of specified percentage divergencies can be determined analytically.<sup>26</sup>

<sup>26</sup> The following symbols are defined for this analysis:  $S$  = sales,  $C$  = costs,  $P = S - C$  = net profits,  $ps = P/S$  = profit rate sales,  $e$  = error in costs due to arbitrary allocations, etc., as a percentage of costs,  $C$ ;  $ps'$  = reported profit rate on sales when costs are over- or understated by  $e$ , and  $ps' = (S - C - Ce)/S = ps - e + eps$ . Since net profits/assets ( $pa$ ) is equal to net profits/sales ( $ps$ ) times sales over assets and neither sales nor assets are assumed to be affected by errors in costs, the effect on  $pa$  is the same as the effect on  $ps$ .

Table 2 shows the reported profit rate on sales or assets that results from a given percentage error in costs, compared to the "actual" profit rates. Divergencies due to nonmarket priced transfers and arbitrary allocations and mismeasurements of assets are ignored; Fisher and McGowan show that divergencies of accounting from economic depreciation alone is sufficient to make the accounting rate of return diverge significantly from the economic rate of return on assets. Even with these potentially important measurement errors implicitly assumed to be zero, the effect of an error in measured cost as small as  $\pm .10$  changes an actual profit rate of .15 to .07 or .24. A  $\pm .20$  error results in a stated profit rate of  $-.04$  or .32 rather than .15. Consequently, it seems evident that data measurement problems certainly might swamp inferences that otherwise might be drawn from the reported numbers.

2. *The Usefulness of Company Accounting Data.* From the analysis presented above, it might seem that accounting data are almost useless. In this event, one might ask, why do companies spend great sums to record, analyze, and report such data? The reason is that even if these data do not provide good measures of economic market values that are appropriate to other issues, they are useful to

The effect on the profit rate on sales or assets also can be expressed as an amount,  $x'$ , rather than as a percentage. In this event, the equation would be  $e(ps - 1) = x'$ .

company managers and outside observers of company activities for several very important purposes. These include internal control for managers and audits of company affairs for investors. These purposes do not require numbers that reflect economic values; rather numbers that reflect market transactions and responsibility for resources are required. When the reported figures are attested to by a certified public accountant, affirmation is made that an audit was conducted by an independent expert who verified the figures and tested the adequacy of the company's financial control system. The accounting numbers also can be evaluated in the context of the manager's or analyst's knowledge about the company's activities and past performance. Thus, the user of the numbers can formally or informally adjust them to account for idiosyncracies in the reported accounting procedures, their company-specific deviations from economic measures, and their relevance to particular questions being asked. (For example, what is the probability that the company can repay debt as promised?) Furthermore, changes in the reported numbers over time, within the context of the past and expected future economic conditions in which the company operates, may provide the user with useful information.

Accounting data also are used for decisions made within the firm, such as output and price decisions, and evaluations of product and managerial performance. For these purposes, the data rarely are used alone. Output and price decisions require estimates of opportunity costs, which often are available only from market data and special studies. Accounting measures of performance usually are evaluated within the context of the environment in which products are produced and managers operate. Thus the accounting numbers rarely are used alone or without specific knowledge of their limitations.

On the other hand, large-scale data bases, such as the FTC's *LB* program set, do not (and, probably even at very high cost, cannot) include much (if any) of the qualifying information available to company managers. When these data cannot be identified by firm

name (as is legally required for the FTC *LB* and Census programs), firm-specific information from other sources cannot be brought in. Thus the usefulness of the data for the FTC's announced purposes and for structure-performance studies generally is, at best, doubtful.

### C. *The Contamination and Completeness of LBs*

Companies normally do not record data according to the FTC-SIC-defined industries. To make the program feasible, the FTC permits the respondent companies to assign to a single *LB* all the data recorded for each "basic component".<sup>27</sup> Consequently, data properly belonging in one *LB* are reported in a different *LB*, resulting in under- and over-statements.

The FTC has sought to measure the extent by which reported sales data are contaminated and complete with a "specialization ratio" and a "coverage ratio." The specialization ratio is the ratio of the sales belonging in an *LB* category ("primary sales") to the sum of these sales and those included in this category but properly classified in another category ("secondary sales"). These ratios averaged .97 (weighted by total revenues) for the 1975 data; none is permitted to be below .85.<sup>28</sup> Thus, *revenue* contamination does not appear to be serious, except for relatively few industries. The coverage ratio is calculated as the ratio of primary sales reported in a category to the total sales of those products reported in all other categories. The weighted average over industries is .97 for the 1975 data. But 29 percent of the industries have percentages between .80 and .91, and 14 percent have

<sup>27</sup>A basic component is defined by the *LB* program as the organizational subunit which reports data. A corporate respondent typically has a number of components, and in many cases they cover products classified by the FTC in different *LBs*.

<sup>28</sup>If at least 85 percent of a component's sales is not accounted for by products classified in the same *LB*, the component must be further refined until the 85 percent rule is met. The ratios are described and the figures reported in FTC (1981b, pp. 41-46). The 1974 figures are similar.

percentages below .80. Thus, for a substantial number of industries, this measure shows a fairly high level of incompleteness in the sales reported.

Of much greater significance is the FTC's candid acknowledgement that "whether there is a significant relation between the degree of sales contamination for other variables [for example, expenses and assets] has not been established" (1981b, p. 43). Nor can it be established with the reported data, since the FTC's 85 percent specialization rules do not apply to expenses and assets, which need not and often do not vary directly with sales.<sup>29</sup> Hence, there is no way of knowing how serious the contamination is for such measures of performance as profit/sales and profit/assets.

Nor are these the only types of contamination. The FTC permits vertically integrated companies to combine data that other companies report as individual lines of business into a single *LB*. This contamination is permitted if more than 50 percent of an upstream (earlier processed) *LB*'s net operating revenues come from transfers to a downstream (later processed or sold) *LB*. A similar combination and contamination of data is permitted for backward transfers (for example, wholesaling combined with manufacturing). As a consequence of these permitted contaminations, additional overstatements and understatements of *LB* data will result.<sup>30</sup>

#### D. *The Stability and Reliability of the FTC's LB Data*

Although there is no way to determine how badly the FTC's *LB* data misrepresent economic market values or are contaminated among *LB*s with respect to expenses and assets, some insight into their validity may be provided by considering the temporal stability of the two key ratios—profit/assets and profit/sales. (Profit is defined as net operating income before interest and taxes.)

Academic researchers, industry analysts, and company managers are invited to use the ratios published in the FTC's *Statistical Reports* (1981a,b; 1982). The ratios reported are aggregates of company *LB*s. Thus differences among companies operating in the same FTC-defined industry are averaged. Nevertheless, the reported industry ratios vary considerably among the three (complete) years for which the data have been published—1974, 1975, and 1976. For example, for the ten median ranked industries in 1976, profit/assets for 1976 compared to 1975 changed by a mean absolute of 37.4 percent, and from 1974 to 1975 by 23.0 percent. The mean absolute change in profit/sales of these industries from 1975 to 1976 is 30.2 percent, and 23.3 percent from 1974 to 1975. The changes among years of the ten highest- and lowest-ranked industries are roughly similar. It is possible that changes in the environment in which the reporting companies operated or changes in the sample caused these differences. The errors of measurement described above also could be responsible for the differences between years.

The data reveal an even greater range with respect to the individual company *LB*s that are aggregated to obtain the industry averages and that are used as the dependent variables for several of the statistical studies discussed below. Excluded from these 1975 data are 889 company *LB*s that were not present in the previous year, had nonpositive assets, or were in SICs described as "miscellaneous" and "not classified elsewhere." Profit/sales (profit equals sales less traceable expenses) averages 7.9 percent (Stephen Martin, 1981a, Table 1, 1975 data, 2297 observations). The standard deviation is 14.1 percent, and the range is from less than -125.8 percent to over 58.9 percent (these are averages of the six lowest and highest percentages). Of these, 16.4 percent are negative and 3.0 percent indicate profitability rates of more than 30 percent.<sup>31</sup> Profit/traceable assets averages 23.6 percent (Martin,

<sup>29</sup>See Bock and my article (1975) for illustrations.

<sup>30</sup>For more on this issue see Bock and my article (1975).

<sup>31</sup>Profitability measured as sales less traceable and nontraceable expenses divided by sales shows a similarly wide range (Ravenscraft, 1982, Table A1).



1981b, Table II, PRA575, 1975 data, 2297 observations). The standard deviation is 170.0 percent and the range is from less than -227.8 percent to more than 2399.6 percent (averages of the six extreme values). Other definitions of assets show similarly wide variations.

If these data reflect the profit rates actually experienced by the reporting companies, it would imply very limited entry and exit into a substantial number of markets. Or, there could be substantial annual random variation in returns, which implies that the data measurement period should be longer than a year. Alternatively, the range of profit rates reported and the number of *LB*s with negative and very high positive profit rates could reflect the effects of accounting practices and allocations to FTC-defined industries. This effect would be exacerbated if the reporting company managers fear that their data will be used against them in anti-trust proceedings, despite the FTC's assurances to the contrary. If so, they can avoid reporting profits that antitrust enforcers might consider excessive or predatory in lines of business the authorities might consider to be suspect. Since the reported numbers must be reconciled with the company's total sales, expenses, and assets, the understatement of profits and profit rates in one *LB* results in an overstatement in the other *LB*s. As a consequence, the FTC anti-trust enforcers would be misled and the data would be subject to even more noise and biases.

### III. Analysis of Economic Studies Based on the FTC's *LB* Data

Though I believe that the biases and shortcomings of reported company accounting data and the FTC's *LB* data delineated above are overwhelming, there are those who strongly believe that these data are valuable if adjusted properly and used effectively. Therefore, the extent to which the reported studies prepared by the FTC's economics staff and consultants can fulfill the purposes specified by the agency are considered explicitly in this section. The working paper

status<sup>32</sup> of the studies reviewed should not shield them from public scrutiny, for several reasons. The *LB* program was proposed 14 years ago—in 1970. The costs of the program on the respondents has been considerable.<sup>33</sup> The nature of the data collected has been known since 1974. Hence, it is reasonable to expect the FTC's economists to have had well-structured hypotheses to which the data would be applied. Furthermore, it is useful to analyze working papers because they are more likely to present unsatisfactory and anomalous as well as satisfactory and expected findings than are published papers.<sup>34</sup> In addition, working papers often contain informative descriptions of the data that are not included in the shortened published revisions.<sup>35</sup> Finally, in this case, working papers can influence policy decisions.

The papers are grouped according to the four major study areas toward which the data appear to have been directed: economies of scale; structure-performance; research and development; and advertising and other selling expenses. All of the papers using *LB* data made available by the FTC through

<sup>32</sup>Several of them have been published, as noted in the references.

<sup>33</sup>See my article (1979, pp. 110–14) for an estimate of the prospective costs. Net of benefits (which were considered to be zero), the present value of the program's cost in 1974 dollars was calculated to be \$210 million, based on the FTC staff's estimate of \$24,000 per company. This estimate was vigorously disputed by the respondents as being greatly underestimated. See my article (1984) for an analysis of the FTC's and the respondents' calculations of the costs of the *LB* program, which concludes that the FTC staff grossly misestimated the costs.

<sup>34</sup>Unfortunately for the sake of knowledge, papers tend not to be published (or even submitted) unless they reveal statistically significant results that are in accord with accepted hypotheses. For example, Weiss reveals: "An empirical test by Hall and me [Marshall Hall and Weiss, 1967] apparently yielded a mild but significant positive relationship between firm size and profit rates among the few hundred largest industrial firms—a relationship which would have implied a capital-requirements barrier. I am more skeptical now because in subsequent unpublished work I have often found the relationship not to be significant" (1979, p. 1122).

<sup>35</sup>In particular, the published version of Martin (1983) does not include the ranges of the *LB* data, and the published version of Ravenscraft (1983) does not include information on the inter-range distribution.

the date when the *LB* program was to have been evaluated (March 1982) are considered so that readers would not think I had chosen only the poor papers.

#### A. *Economies of Scale*

Absent analysis of scale economies, there is no way of determining whether a measured positive relationship between profit and concentration was due to some sort of undesirable competitive behavior or to lower costs.<sup>36</sup> Obviously, the policy implications of these alternative explanations are diametrically opposed. Unfortunately, neither the paper reviewed next (the only one to address this question), nor any other study based on the *LB* data, can provide a meaningful measurement of economies of scale.

1. In "*Economies of Scale, Concentration, and Collusion*," Dennis Mueller (1980) states his objective as follows: "What we seek to explore in this [61 page] study, therefore, is the extent to which the efficiency-concentration-collusion interrelationships can be disentangled using data as disaggregated as the *LB* data are" (p. 5). Mueller attempted to estimate a Cobb-Douglas production function by regressing value-added (sales less materials) on payrolls, net plant and equipment/payrolls, and (in some equations) advertising/payrolls. He had to "assume that labor and capital stock are somehow pre-determined and exogenously given for any cross-section year" (p. 8). As an alternative measure of economies of scale, he regressed total cost on sales. The equations were run for each *LB* industry, the observations being company *LBs*, 1974 data. The estimated alternative measures of scale economies are weakly and negatively correlated, which is an indication that the models are poorly spec-

ified. But a much more important criticism is that neither of these forms could be meaningfully specified. The dependent variables, sales less materials and total costs, are determined by the interaction of supply and demand (and, importantly, by the companies' accounting systems). Therefore, even if the numbers used were meaningful measures of economic market values, there is no basis for determining whether the regressions trace out cost curves, demand curves, or the interaction of individual demand and cost curves. Thus there is no valid way to interpret the estimated coefficients.

Despite this crippling limitation, Mueller correlates the mislabeled "scale economies estimates" with several measures of concentration, stating: "If scale economies account for high concentration we expect a positive correlation between [the measures of scale economies and] the various concentration measures" (p. 25). Since the measures could reflect demand as well as supply conditions or simply bad data and poorly structured models, the small and generally negative correlations that he finds cannot be meaningfully interpreted.

In the penultimate section of his paper (ch. 4), Mueller attempts to distinguish collusion and economies of scale by assuming that all firms use the same technology, have the same cost function and the same products, and charge the same price. Given the broad FTC-SIC definitions of industries, these assumptions surely do not apply to the *LB* data set. Nevertheless, after some mathematical manipulations, he concludes: "Thus, while efficiency and collusion factors affect the pattern of costs and sales in a predictable way, their effects cannot be separated without imposing further restrictions and looking at more data than observations on firms within a single line of business" (p. 33). He then imposes several additional assumptions that, in effect, specify all firms in each *LB* industry as homogeneous with respect to economies of scale, degree of collusion, and elasticity of demand. After further manipulations and assumptions about the values of various parameters, Mueller concludes: "Attempts to estimate [the key equa-

<sup>36</sup>Even then economies of scope are ignored due to the separation of company data into FTC-designated lines of business. Furthermore, other efficiencies could be responsible for higher profit rates, as Peltzman suggests. Or efficiencies in the development and/or marketing of new products could be responsible, as Scherer (1979b) suggests.

tions] yielded disappointing results. Convergence was seldom achieved and even when achieved, the asymptotic standard errors often exceeded the regression coefficients. Moreover, whether or not congruence was achieved, the implications of the coefficients obtained were found to vary with the choice of deflator" (p. 47).

Thus the *LB* data are not even amenable to estimating what might be labeled (incorrectly) as economies of scale, without which the structure-performance studies cannot be used for policy purposes. Mueller's study is useful because it illustrates that the unreliability and inherent limitations of the *LB* data severely frustrate even a well-motivated investigator.

### B. Structure-Performance Studies

1. In "Structure-Profit Relationships at the Line of Business and Industry Levels," Ravenscraft (1982) inquires "whether profits rise with industry concentration when other structural variables, such as market share, are appropriately held constant. Also, what economic phenomena underlie the observed profit-market share associations?" (p. 17). Profit is defined as sales less direct (traceable) and allocated expenses divided by sales, as taken from the 1975 *LB* data for 3186 company *LB*s. Profit at the industry level is a similarly measured number derived from the 1975 *Annual Survey of Manufacturers* for 258 industries.<sup>37</sup> Those dependent variables are regressed on the fourteen variables listed in Table 3 plus nine additional variables for the company *LB* analysis. The industry regressions exclude the variables that relate only to company *LB*s.

Ravenscraft's study is seriously marred by the conceptual and empirical shortcomings of the dependent variable, profit/sales, which

also is used in several of the papers reviewed below as well as scores of others.<sup>38</sup> This ratio is justified as a proxy for the price-cost margin (price less marginal cost over price)—the Lerner Index. However, as Stanley Liebowitz (1984) shows, the Lerner Index is not a conceptually useful measure of monopoly or collusive pricing on which the antitrust authorities might base their actions (an emphasized goal of the FTC's *LB* program). At most, it only measures the marginal elasticity of demand; it does not measure the social cost of misallocated resources or even the amount of profits or wealth garnered from collusive pricing. Furthermore, differences in the price-cost margin among products or over time is a function of differences in marginal costs and exogenous shifts in demand.<sup>39</sup> Importantly for empirical work, as Stanley Ornstein (1975) demonstrates, the profit/sales ratio is not even a valid proxy for the Lerner Index. Recorded and allocated accounting expenses are not the same as economic costs on which the Lerner Index is based.<sup>40</sup> In addition, profit/

<sup>38</sup> Profits/assets is not used in any of the FTC-sponsored studies, except as follows. Ravenscraft states that: "With the ratio of operating income to *assets* as the dependent variable, only three conclusions change" (1982, p. 11, fn. 11). He does not otherwise mention this alternative dependent variable. Martin (1981b) regressed profits divided by six versions of assets as dependent variables. The coefficients of the independent variables differ considerably in significance and sometimes in sign from those of the same variables when profit/sales is the dependent variable. Furthermore, the coefficients differ considerably among the alternative measures of assets, often much more than can be accounted for by scale differences. For example, the Herfindahl index coefficient is statistically significant for only one version of the profits/assets, and three of the six coefficients of market size are significant.

<sup>39</sup> See Liebowitz (1984) for a thorough explication.

<sup>40</sup> Liebowitz (1982) shows empirically that the empirical price-cost margin is not a good proxy for either accounting profit over assets or equity, or a constructed measure of economic profits over sales (which ignores the accounting biases). Martin dismisses Liebowitz's work as follows:

Liebowitz (1982) criticizes the Census price-cost margin by comparison with Internal Revenue Service data. Scherer identifies two major problems with IRS data: the assignment of entire firms to a single industry (1980, p. 270) and the impact of accounting rules which are

<sup>37</sup> The regressions were replicated with 1974 and 1976 data. With a few exceptions, the coefficients are not significantly different among the samples, and almost all were statistically significant at the .10 level, in part because, as Ravenscraft states, "variables insignificant at the 10 percent level were eliminated by a stepwise procedure" (1982, p. 8).

sales differ among companies according to their degree of vertical integration and the extent to which they employ capital (since the factor price of capital is not included as an expense). Including assets/sales (or assets) as an independent variable to account for the bias is likely to lead to inefficient and biased estimators because the accounting number for assets usually is a very poor measure of the economic value of capital.<sup>41</sup> Indeed, in Ravenscraft's regressions (as in the other researchers' regressions), the coefficient of assets/sales is inexplicably significantly negative. It is likely that this coefficient, and probably others, is due to accounting-based biases.<sup>42</sup> Finally, the rela-

---

followed for tax purposes only (1980, p. 272). As noted above... Liebowitz corrects for the first problem (1982, pp. 238-39, footnote 22) [sic, fn. 21]. He recognizes the second (1982, p. 238, footnote 20), and assumes that it can be ignored. There is no reason to think this is the case, which invalidates his conclusion.

[1983, p. 32, fn. 20; emphasis added]

This criticism is curious, since it appears in a section of Martin's paper headed "The Lerner Index and Accounting Data," in which none of the measurement problems delineated in Section III above are mentioned. Furthermore, the accounting biases in the income tax data, about which Martin expresses concern, are no more serious than the biases in the FTC data that he uses, since the value weighting of the IRS sample overwhelms the additional biases imparted by small firms. Indeed, the IRS data are preferable, since they are not subject to the allocation biases inherent in the FTC *LB* data. (Also see Scherer, 1979b, pp. 200-05, for a demonstration of the biases introduced into industry numbers by the incorrect assumption of homogeneity.)

<sup>41</sup>For example, oil and gas producers severely under-value their major asset, oil and gas reserves, because they do not restate their assets at market values. Worse yet, large companies use successful efforts accounting, wherein only the exploration and development expenses of successful wells are capitalized. Their cost of sales, then, is understated if they use owned oil. But if they use imported oil (which, if purchased, is valued at market), these expenses are higher. If they are decreasing their exploration, drilling, and development activities, their expenses will be lower. Thus, how can the level of their profit rate on sales or, say, the coefficient of, assets/sales, be interpreted? Similar analyses could be made about any company or industry about which one is knowledgeable.

<sup>42</sup>An additional indication of such biases is given in Table 3, which compares the coefficients estimated by Ravenscraft with those estimated by Martin (1981a), who used similar data. (Note that most differences in scaling among the papers were adjusted for.)

tionship between profits/sales and concentration (or market share) should be specified as log linear if it is to serve as a proxy from profits/capital. Ornstein shows that, if this is not done (as it was not in the studies reviewed), statistically significant coefficients incorrectly result.

Aside from (or perhaps because of) the biased numbers used for the analysis, the meaning of the coefficients is difficult to assess. For the company *LB* regression, Ravenscraft finds that "Statistically, the most important variables are the positive effect of higher capacity utilization and industry growth, with the positive effect of market share running a close third" (1982, p. 11). Capacity utilization is measured as the smaller of "one" or the company's *LB* 1975 sales/1974 sales; industry growth is the ratio of the value of 1976 to 1972 shipments. Thus, the greater the increase in company *LB* sales or in industry shipments, the greater the increase in measured profit per dollar of sales. Does margin increase because expenses are less than proportionately variable? Or, were companies in expanding industries able to increase their selling prices more? Or, were industries with higher profit margins on sales able to grow faster?

Interpretation of reported market share is more important because this presumably bears on the profit-structure relationship in which Ravenscraft and others are interested. Ravenscraft expects a positive association of market share with profits for three reasons: "First, *LBs* with a large market share may have higher quality products or market power. Second, *LBs* with larger shares may be more efficient because of scale economies or because efficient *LBs* tend to grow more rapidly. Third, *LBs* with large market shares may be more innovative, or better able to develop innovations" (1982, pp. 2-3). Clearly, these alternative explanations not only have diametrically opposed policy implications (higher market power vs. other explanations), but they cannot be disentangled. In addition, if firms with larger market shares also have been in business longer and have relatively greater investments in fixed assets over an inflationary period, they will have lower depreciation expenses and possibly

TABLE 3—COMPARISON OF COEFFICIENTS ESTIMATED IN FOUR STUDIES USING FTC'S *LB* DATA<sup>a</sup>

Variables	1974 Data		1975 Data	
	Ravenscraft	Martin	Weiss-Pascoe	Scott
Market Share	14.76 (5.51)	10.09 (2.98)	-0.20 (6.47)	0.11 (1.80)
4-Firm Concentration Ratio	-0.02 <sup>b</sup> (1.77)	-0.11 (5.77)		
Diversification (firm)	-1.43 <sup>b</sup> (1.65)	-1.99 (1.30)		
Minimum Efficient Scale (industry)	0.18 <sup>b</sup> (2.05)	0.39 (2.97)		
Exports/Shipments (industry)	6.51 (1.72)	11.25 (2.32)		
Imports/Shipments (industry)	-4.01 (2.23)	-2.05 (0.81)		
Buyer Concentration Index (industry)	5.52 (4.48)	0.08 (3.60)		
Supplier Concentration Index (industry)	-31.40 <sup>b</sup> (1.39)	-.09 (2.58)		
Buyer Dispersion	-0.46 (0.64)	-6.13 (4.00)		
Supplier Dispersion	-4.59 (2.86)	-7.20 (2.79)		
R&D Expenditures/Sales ( <i>LB</i> )	-0.47 <sup>b</sup> (3.68)	-0.31 (5.62)		
Assets/Sales ( <i>LB</i> )	-2.40 (2.82)	-5.33 (6.75)	-0.032 (5.64)	-0.0082 (0.59)
Assets/Sales (industry)	6.00 (4.98)	1.87 (1.31)		
<u>Similar Variables</u>				
Selling Expense (direct)/Sales ( <i>LB</i> )		-0.30 (6.74)		
Advertising/Sales ( <i>LB</i> )	-0.02 <sup>b</sup> (0.35)		0.123 (1.90)	-0.083 (0.54)
Growth in Demand (shipments)			0.061 (10.76)	0.046 (3.1)
Distance Goods Shipped			-0.000018 (3.07)	-0.000021 (1.2)
Number of Additional Variables	9	24	3	3

Notes: Ravenscraft (1982, Table I, 3186 observations); Martin (1981a, Table 15, corrected for heteroscedasticity, 2297 observations); Weiss and Pascoe (1981, Table 1, equation (4), 3043 observations, scale not given); Scott (1981, Table 7, equation 7-2, 480 observations, scale not given).

<sup>a</sup>Absolute *t*-statistics are shown in parentheses.

<sup>b</sup>Rescaled to Martin's magnitudes.

lower labor and other expenses (as capital is substituted). Consequently, higher market shares will be associated with higher profits, *ceteris paribus*. Or, the reporting companies could have allocated relatively less in common costs to the areas where they have greater market shares. Or, possibly transfers to these units were made at less than market prices. Without more information, implica-

tions should not be drawn from the coefficients measured.

Ravenscraft attempts to analyze a presumed market share (*MS*) relationship to profit/sales by regressing profit/sales on *MS* and *MS* times the other *LB* variables. He finds a negative coefficient for advertising/sales, a positive coefficient for industry advertising/sales, and a positive co-

efficient for the product of advertising and *MS*. He estimated that the *LB* companies made 47 percent of the sales in their industries, on the average, and calculated that profit/sales begins to increase when *MS* rises above 4 percent (p. 17). Assuming that these numbers are valid economic measures, should they be interpreted as meaning that firms with more than trivial market shares charge higher prices, offer higher quality products, economize more on other selling expenses, have a greater investment in capital, or what? Interpretation of the other measured relationships are subject to similar uncertainties.

Thus, despite Ravenscraft's imaginative econometric work, little can be learned, with one exception. He finds that when the company *LB* data are aggregated into industries, the results are different. The coefficient of the four-firm concentration ratio changes from significantly negative to significantly positive (probably, he notes, because company *LB* market share is not included in the industry regression). Five of the coefficients of the fifteen other common variables change from significance to insignificance (three also change sign), and all but two of the balance increase in magnitude by from two to five times. It appears, therefore, that analyses using industry data may have produced misleading results. Or do the regressions using company *LB* data produce misleading results? Or are the companies' allocations to FTC-defined *LBs* causing the differences? Or are the changes in the coefficients caused by differences in the independent variables specified in the company *LB* and in the industry regressions? At the least, these alternatives should lead one to wonder about the usefulness of the output of such exercises for public policy or business decisions.

2. "Market, Firm, and Economic Performance" (Martin, 1981a; 1983). Martin attempts to analyze the determinants of the price-cost margins of company *LBs*, where the profits/sales dependent variable was measured as reported sales (including transfers) less traceable expenses only divided by sales. Martin (1981a) specified thirty-two independent variables. Though he considers seventeen of these variables to be

endogenous, he estimated simultaneous equations only for the *LB* market share, firm market share, and *LB* selling efforts variables. The two-stage least squares estimates were corrected for heteroscedasticity; the corrected and uncorrected coefficients are presented. The observations are 2297 company *LBs* for 1975. Excluded are *LBs* that were not surveyed in 1974, and observations with nonpositive values for assets (since assets were converted to logarithms).

In Martin's profit/sales regression (1983), twenty of the twenty-seven coefficients reported are statistically significant at the .05 level or better, and four were at the .10 level.<sup>43</sup> Among the findings that might be of interest to the regulatory authorities, Martin finds a statistically significant (.01 level) positive coefficient for *LB* market share and a statistically significant negative coefficient for the Herfindahl concentration index. He concludes that "the market share coefficient may reflect either market power at the *LB* level, or the realization of scale economies, or both" (p. 48). Thus, without further information (which is not obtainable from the *LB* data), this finding is of little public policy value.<sup>44</sup>

Assets/sales, which was supposed to account for the unrecorded cost of capital em-

<sup>43</sup>He does not state (as does Ravenscraft) whether variables were eliminated with a stepwise procedure.

<sup>44</sup>In his earlier paper, Martin (1981a) attributed the unexpected negative coefficient of the Herfindahl index to a variety of possibilities (for example, suboptimal or excess capacity, X-inefficient, inflation, restrictions on exit following restrictions on entry, and market power expressed as reduced risk). He also recognized that the profitability measure (the dependent variable) does not reflect profit, but the profit rate on sales; hence, he acknowledges, the negative sign "may reflect a more than proportionate impact of market power on sales as compared to profit" (p. 33). However, in the published version, Martin states: "Although a number of ad hoc explanations could be offered for this result, I will not do so" (1983, p. 40). But in his conclusion he states: "The most likely explanation for this result is that oligopolistic coordination is less effective in recession years, while the very barriers to entry which engender market concentration impede exit in the short run" (p. 48). In neither the FTC nor the published version did he consider the possibility that the significant negative coefficient also might indicate misspecification of the model or data that do not reflect economic market values and impart biases.

ployed, has a negative, statistically significant (at the .01 level, with the second highest asymptotic *t*-statistic) coefficient.<sup>45</sup> This unexpected result is consistent with the "bad data" hypothesis, since the sign of this variable should be positive.

Two additional facts support the inappropriate model or bad data hypotheses. One is the values of the dependent variable, sales less traceable expenses/sales, which range from at least -125.8 percent to more than 58.9 percent (each are averages of six extreme values). Negative returns are present in 16.5 percent of Martin's observations (1981a; these data are not reported in the 1983 monograph). (Recall that 889 observations were screened out because they were new or labeled as miscellaneous.) As noted above, it seems very doubtful that these numbers could represent meaningful equilibrium economic profit rates. Second, as shown in Table 3, many of the coefficients (adjusted for differences in scaling) and *t*-statistics reported by Martin differ considerably from those reported by Ravenscraft (1982), even though they based their analyses on subsets of the same 1975 *LB* data. (Their dependent variables are somewhat different, since Ravenscraft deducted allocated expenses from sales while Martin did not; but they used thirteen independent variables that are almost identical.) These differences make descriptive statements about the profit-structure relationship of doubtful validity.

The possibility that biases in the data might have been responsible for his results is not recognized by Martin. In his published version (1983; a monograph, not constrained by extreme concerns for space), he hardly mentions biases in the data. Though he has a section partially labeled "accounting data," this contains only a conceptual criticism of Fisher and McGowan and some sentences stating that though "accounting measures...are likely to be...poor measures of the economic value of the capital stock [,]

...[t]his point is well known. ...[T]his measurement error is a serious problem only if it is systematically related to market structure. It is not obvious that this is the case" (p. 32). But, as is noted above in the discussion of Ravenscraft's paper (1982), such systematic biases could be present.

Furthermore, many of Martin's other findings can be attributed to biases in the data. For example, he reports that "non-advertising sales efforts at the firm level significantly increase *LB* profitability. A firm-level corporate marketing program is apparently a valuable asset, which increases profitability" (1983, p. 41). But, since Martin used *LB* profits before company-unallocated expenses, this finding also could simply reflect differences in accounting practices among companies, where those that allocate less (for example, centralized firms) have, as a consequence, higher measured *LB* profit rates and higher selling expenses at the firm level, *ceteris paribus*. Martin also concludes that "*LBs* which invest heavily in *R&D* are less profitable than other *LBs*, all else equal. This suggests that heavy *R&D* expenditures serve as an index of the intensity of conventional price competition, which apparently serves as a spur for research and development" (p. 42). But, since current *R&D* expenditures are called expenses, the finding might simply reflect accounting practice, and little else. On the other hand, firms that successfully invested in *R&D* in the past do not record the depreciation of this investment; hence their recorded profits are higher, *ceteris paribus*. Just how this might be related to market shares is not clear. In addition, Martin concludes that "a large firm-level *R&D* program enhances *LB* profitability" (1983, p. 46). But again this simply could reflect the interaction of Martin's measure of profitability and accounting allocation practices. Were there sufficient space, similar explanations of many of Martin's other findings could be provided.

3. "Some Early Results on the Concentration-Profits Relationship from the FTC's Line of Business Data" (Weiss and George Pascoe, 1981). The authors want to test Demsetz's (1974) hypothesis that concentration-related economies of scale (efficiency),

<sup>45</sup> Martin (1981b) explored this result more fully by running regressions with six alternative definitions of assets. (Assets defined as all except inventories fit the data best.) The results were substantially unchanged.

superior products or management, or luck is responsible for the often reported positive relationship between concentration and profits. To this end, Weiss and Pascoe regressed profits/sales on the variables listed in Table 3 plus an adjusted four-firm concentration ratio (adjusted "to correct for non-competing sub-products, inter-industry competition, local or regional markets, and imports" p. 2), an interactive concentration ratio equal to the concentration ratio times a ratio of consumer demand to total demand, imports as a percentage of domestic output plus imports less exports, and *LB* market share. The regressions were run with over 3000 observations, each a company *LB*, with 1974 and 1975 data separately. The *LB*s that were new (not reported in 1974) and discontinued (not reported in 1975 or 1976) were excluded, as were extreme outliers. The reported  $R^2$ s range between .04 and .07. The coefficients of market share are statistically significant ( $t$ -statistics over 5.6) and positive. Those for concentration are insignificant, except for the 1975 data when market share is excluded. The coefficients of the interactive concentration ratio is significant and negative in 1974 and not significant (though negative) in 1975. Thus market share dominates which, Weiss-Pascoe state, "provides some support for the Demsetz-Manke hypothesis, but it is not as compelling as it might have been if the relationship of margin with concentration excluding market share had been stronger" (p. 8).<sup>46</sup>

But, as noted above, profit per dollar of sales of companies in different SIC-defined industries is not a relevant measure, particularly since the coefficient of assets/sales, which is included to correct for the exclusion of the factor cost of capital, is significantly ( $t$ -statistic over 5.6) negative. In an effort to explain this result, they regress assets/sales on the other independent variables. They

offer some speculations about the results, but state: "These are at most hypotheses to be tested on other data and, hopefully, with some theoretical basis" (p. 11). I suggest that the observed and unexplained relationship between profit/sales and assets/sales is a measure of the divergence between economic market values and the accounting numbers used.

4. In "*Multimarket Contact and Economic Performance*," John Scott (1981) is concerned with measuring whether the presence of the same firms in several markets is associated with higher profits and, if so, whether these higher profits result from coordinated behavior, lower costs, or both. To answer these questions, Scott constructed a "new method of assessing multimarket contact" (p. 3). For each of twenty-four lines of business (1974 data) he paired 492 "competing" firms (246 pairs), based on the unverified assumption that firms with sales assigned to the same FTC-defined line of business necessarily were competitors. Then he recorded the number of other lines of business each reported and the number they reported in common. This "frequency of contact" was compared to the number expected if common *LB*s were randomly determined. He finds "significant multimarket contact among sellers" (p. 8), as evidenced from a greater number of observed common *LB*s (37 percent of the pairs) than are expected by chance (.10 level). But considering that firms are more likely than not to produce related products, this result is not surprising nor particularly significant. Also, since fewer than the expected number of common *LB*s (.10 level) were found for 31 percent of the pairs, the distribution is strongly bimodal. This result seems more surprising.

The fifty-one pairs of firms that had more contacts than expected by random selection were subjected to further tests. Scott finds "significantly less advertising and *R&D* intensities [ratios to sales] for lines of business where significant contact occurred" (p. 13). Apparently, firms that operated in related *LB*s experienced economies for some reason. In another test, Scott regressed operating profit/sales on the variables listed in Table 3

<sup>46</sup>They also find that "distance shipped" (a measure of geographic market share) is statistically significant (and negative). This is curious since the variable is an SIC industry-wide index based on 1963 data, which should be only tenuously related to the 1974 and 1975 company *LB* dependent variables.



plus a measure of industry minimum efficient scale, a dummy variable for the probability of greater multimarket contact above the sample median, and a dummy variable for a four-firm concentration ratio above the sample median. The product of the dummy variables also was included as a variable to measure the interaction of higher contact and higher concentration. The 492 selected firms (246 pairs) are the observations. With respect to his hypothesis, Scott finds significantly (.10 level) greater profit/sales that is about 3 percent higher for the higher-contact/higher-concentration variable. But whether this is due to greater efficiencies, barriers to mobility, collusion, accounting and classification biases, a relationship between vertical integration and number of "contacts," or some other factor cannot be determined from the analysis. Furthermore, as Table 3 shows, there are considerable differences between several of the *t*-statistics and signs of the coefficients Scott estimates and those estimated by Weiss and Pascoe, which renders the findings suspect.<sup>47</sup>

### C. Research and Development

1. In "Using Linked Patent and R&D Data to Measure Inter-Industry Technology Flows," Scherer (1981a) describes the procedures used in gathering data on 15,062 patents issued to 443 large industrial companies during a ten-month period (June 1976 through March 1977) and in assigning each patent to one or more lines of business. The companies' 1974 R&D expenditures per the FTC's *LB* reports then were associated with the inventions. The paper is very clear and honest in describing the great difficulties that were experienced in making the required associations. Though he was most imaginative in attempting to solve the problems, Scherer's expression of them makes it clear that they were "solved" only by making some heroic assumptions that severely compromise the meaningfulness of the resulting numbers.

Aside from questions of the meaningfulness of the data, the essential question is why the analysis was undertaken in the first place. In his introduction, Scherer states: "The motivation for developing these new data was straight-forward. During the 1970s... productivity growth [declined]. ...Beginning in the late 1960s,...privately-financed real R&D [similarly declined]. ...The key questions remain, what quantitative links exist between R&D and productivity growth, and did the parameters of any such relationships shift between the 1960s and the 1970s?" (pp. 1-2). Even if there were no questions about the validity of the data or the presumed decline in productivity, the matrix associating the number of patents issued and expenditures on R&D by *LB* that is presented at the close of the paper cannot answer this question since the R&D expenditures are not related to changes in productivity. Indeed, at best, all that is learned is that expenditures on R&D were made in some FTC-defined industries and that these expenditures *may* have been related to the output of these industries and other industries.

2. "The Propensity to Patent," Scherer (1980) uses the data just described because "they are a rich source of quantitative and qualitative information on technological change" (p. 1), and, to a lesser degree, on monopoly power. To these ends, Scherer regressed the number of patents issued to a company *LB* from June 1976 through March 1977 on eight technology class dummy variables, five variables measuring the type and scope of use, two variables designed to correct for some *LB* data reporting characteristics, the percentage of federal funding to total R&D expenditures, and a variable described only as "compositions of matter." All of the variables are multiplied by R&D expenditures and a constant term was added. The regressions were run with 1819 observations.

How this specification and the reported findings speak to the questions that presumably motivated the study is not clear. If the question asked is whether greater expenditures on research and development are positively related to a greater number of patents issued, an answer would hardly require a

<sup>47</sup>Differences in the coefficients' magnitudes could be due to the scaling of the variables; the means are not reported.

rigorous empirical study. It seems obvious that if more is spent on research and development, more patents will be applied for and granted. Another unremarkable major finding is that there are differences among industry groups. None of the empirical data provide "information on technological change." Monopoly power is measured by the concentration ratio, which was used to test the hypothesis that firms in more concentrated industries use patents as a means of securing their positions. The statistically significant, negatively signed coefficient would seem to disprove this hypothesis. Or, a critic of business behavior could argue that more highly concentrated industries have erected such great barriers to entry that they need not engage in further patenting. But Scherer dismisses this finding as being due to the highly concentrated automobile industry, which, he says, has few inventions. Or, given the quality of the data, there could be other biases present, as noted above.

3. In "*Inter-Industry Technology Flows and Productivity Growth*," Scherer (1981b; 1982) says that he "exploits a new, uniquely rich data source to analyze the relationships between research and development (*R&D*) and productivity growth" (1982, p. 627). Using the FTC's 1974 *LB* data, he associated *R&D* expenditures (cleverly, but given the nature of the data, necessarily crudely) with originating and using industries. He then regressed two productivity measures on the 1974 *R&D* expenditures divided by value-added and aggregated into two-digit manufacturing groups.<sup>48</sup> The productivity dependent variables are "total factor productivity growth" (not otherwise described) and sales/labor expense, expressed as changes over 1948–66, 1964–69, or 1973–78. Other analyses present simple correlations of various measures of productivity growth with *R&D* by industry of origin and industry of use,

broken down into "well measured" and "poorly measured" industry aggregates.

Unfortunately, Scherer infers causation from the correlation of *R&D* expenditures and productivity without providing a supportive conceptual or empirical basis. For example, a company could have higher measured productivity growth in a period which yields higher present or expected taxable profits. As a consequence, it might expend more on *R&D* because this is a tax deductible expense. Thus, the higher productivity (and profits) could have "caused" the higher expenditures on *R&D*. Or the company's 1974 expenditures on *R&D* might have been financed from the fruits of past productivity, and directed towards reducing material and capital expenditures and developing new products, rather than towards further increasing labor productivity. In this event, past productivity would be associated with present *R&D* expenditures, but these *R&D* expenditures would not be associated with future measured productivity. Or the regression coefficients simply may reflect mismeasurements. (The productivity measures are very crude.) And the amount expended on *R&D* is unlikely to measure the change in the stock, except in a steady state, at best. In addition, the industry aggregates are rather crudely defined. Therefore the regressions cannot provide evidence on the relationship between productivity and *R&D*, or even on whether research productivity has declined over some time period. Nor is Scherer justified in concluding "that the social returns during the 1970s [on *R&D* investment] appear to have been quite high" (1982, p. 633), since this conclusion is drawn from the coefficients of the regression of the productivity measures on *R&D* expenditures. There is no way of knowing from these data whether there is a causal relationship or even what the social return might have been.

4. "*The Effects of Inter-Firm Cooperation and Economies of Scale on Product Improving Research and Development Expenditures*" (Long, 1981b) begins: "The purpose of this paper is to perform an empirical analysis of the determination of research and development-

<sup>48</sup> The model calls for percentage changes in the stock of *R&D*, capital/labor and materials/labor. The *R&D* expenditure was assumed to measure the first variable; the other two were omitted.

expenditures for the improvement of the quality of manufactured products in the American economy for 1974 and 1975" (p. 1). For this purpose, Long proposed a model of *R&D* behavior and ran several regressions with data from 205 FTC-defined industries from 1974 to 1975 individually. It is difficult to evaluate the findings of this work, since it is incomplete. One thing that was learned is that the relationship between reported *R&D* expenditures/sales and number of patents issued differs considerably among FTC industries and, remarkably, often is negative and differs in sign between years. It is not clear whether this is a function of some underlying structure or of the quality of the reported data. In any event, it seems clear that the stated purpose of the paper—"the determination of research and development expenditures for the improvement of the quality of manufactured products"—neither was nor, given the FTC *LB* data, could have been fulfilled.

#### D. Advertising and Other Selling Expenses

1. In "*The Size of Selling Costs*," Weiss, Martin, and Pascoe (1981) seek "to explain selling expenses in manufacturing industries" (p. 5). They regressed advertising/sales, other selling costs/sales, and total selling costs/sales (separately) on five independent variables. The observations are 260 FTC-defined industries (1975 data). The authors learn that all three measures of selling costs/sales are positive significant functions of consumer demand/total demand, the distance products are shipped, and profit/sales—results that are not surprising. But they also find that selling cost/sales declines for concentration ratios greater than .51 for advertising and .39 for other selling costs. The descriptive or public policy implications of this finding (assuming that it is valid) are unclear. Do companies in more concentrated *LB* industries incur less selling expense/sales because they have achieved efficiencies? Or do they compete less in terms of selling expenses? Or have they spent more on selling in the past and now need not maintain consumer acceptance, while companies in less

concentrated industries have yet to inform (or possibly influence) consumers? These and other plausible hypotheses cannot be distinguished with the FTC *LB* data.

2. In "*Advertising Intensity, Market Share, Concentration and Degree of Cooperation*," Long (1980) wants to "assess the role that advertising plays in several explicit models of industrial organization and to formulate procedures for testing hypotheses which that assessment generates" (p. 1). To this end he developed a model which led him to estimate the following equation for each of 32 consumer goods "industries":  $\text{advertising}_i / \text{sales}_i = B_0 + B_1 \text{ market share of } i\text{th firm}$ , where  $i, \dots, n = \text{firms in an industry}$ . The regressions were run with 1974 company *LB* data and replicated in Long (1981) with 1975 data. While the  $B_0$  estimated are quite stable and generally statistically significant, relatively few of the  $B_1$  are significant, and many are inconsistent in magnitudes and even signs between years. Long concludes: "The predicted relation between advertising intensity and market share shows up clearly in only 25% of the cases examined; ...some evidence concerning the presence and impact of cooperation was produced, but it is not clearly persuasive; ..." and "virtually no evidence concerning economies of scale in advertising can be gleaned from this study, given its assumptions" (1980, p. 28). Whether the model tested has been largely disproved, was inadequately specified, or was unsuccessful due to the poor quality of data used cannot be determined. I suggest that if one reads the SIC descriptions of the *LB* industries listed in Long's Appendix B,<sup>49</sup> considerable weight would be given to the last explanation.

<sup>49</sup>For example, the first five "industries" are the following: meat packing, sausages and other prepared meat products, including bacon, ham, canned meats, and smoked and fresh meats; dairy products excluding fluid milk, including butter, cheese, condensed and evaporated milk, and ice cream; canned specialties, including baby food, baked beans, ethnic foods, health foods, and soup; frozen specialties, including baked goods (except breads), dinners, pizza and waffles; canned, dried, dehydrated, and pickled fruits and vegetables, including preserves, jams, jellies, dehydrated soup mixes, vegetable sauces and seasonings, and salad dressings.

#### IV. Summary and Conclusions

Several factors make aggregates of company accounting-determined data, such as those gathered by the FTC's Line of Business Program, of doubtful value for the purposes of economic analysis. Most important, perhaps, is the fact that the numbers reported, which are derived from the companies' accounting systems, do not reflect economic market values well. The biases caused by the divergencies of accounting numbers from economic values not only are likely to be considerable, but also impart biases to most of the reported numbers, particularly profits and assets. There is reason to expect the errors not to be randomly distributed and to yield measured, though invalid, relationships among the variables that purport to reflect the economic performance of companies in markets.

An additional important error follows from the FTC's use of the SIC categories and from the procedures companies must or can follow in reporting their data. The SIC-designated "industries" yield aggregates that are unlikely to be comprised of products that are substitutes in demand. The FTC's aggregation rules and purposeful reporting by companies that fear the data might be used against them further bias the data, so that it is not clear how well the reported numbers relate even to the industry categories as described in the SIC manual. While the FTC staff made some tests of the extent to which the sales were misallocated, there is no way that they can determine the extent to which expenses and assets were misallocated.

A review of the reported underlying numbers reveals that the variance of such measures of profitability as profit/sales and profit/assets is great. A fairly high proportion of the numbers are negative and the profit rates measured for a number of lines of business is enormously high. It is possible that these data reveal the profit rates actually experienced by the reporting companies and that a year is too short a period for useful measurement. However, I suggest that these data reflect the accounting biases present in the numbers. In either event, it is doubtful whether analyses using these data would yield

valid findings.

The twelve studies done by the FTC staff and consulting economists confirm this conclusion. Even aside from the data problems, the potentially important study on economies of scale is fatally flawed in concept. Without knowledge of economies of scale, the profit-structure studies cannot distinguish the effects of efficiency from the effects of collusion on a measured positive relationship between profitability and market structure. But most of the studies use profit/sales as their measure of profitability. It is not clear at all why profit before interest and taxes per dollar of sales is a measure of "profitability" among companies with different accounting procedures that greatly affect the rates. Nor does this metric, even were the accounting biases somehow adjusted, provide a useful measure for public policy or private resource allocation decisions. But these studies do provide some indication of the mismeasurement of the numbers, in that assets/sales was found to be strongly negatively related to profit/sales, while the reverse should be the situation. The studies that relate current expenditures on research and development to patenting and productivity infer causal relationships when there is no reason to expect them from the nature of the data that are or could be used. The studies concerned with selling expenditures are useful only in providing some descriptive data. Considering the limitations of the relatively well done econometric studies that use the considerably improved FTC *LB* data, it seems likely that other studies using less detailed data and often less sophisticated statistical methods are likely to be worse.

A final question that should be confronted is whether meaningful empirical work on the relationship between structure and performance is feasible. The answer, I believe, is that such work is useful, but not in the way much of it has been conducted. Joe Bain's (1951) often cited paper has led many economists to undertake and publish a very large number of misleading studies. These papers present what purports to be evidence that concentrated markets are associated with abnormally high profits, when it is likely that the positive coefficients were simply the re-

sult of biases in the accounting data used. The consequence not only is a misunderstanding of the working of the U.S. economy, but a possibly misguided antitrust policy that prohibits most horizontal and many other mergers.

If strong and convincing work on the relationships among profits, company size, market share, degree of competition and the like is to be conducted, it must take cognizance of the biases inherent in accounting and other data. It is very doubtful, though, that the required adjustments to the data can be made, particularly when these are aggregated in ways that cannot take account of significant differences among supposedly homogeneous observations. Rather, detailed company and industry studies, which require knowledge of the specifics and dynamics of events and institutions more than of econometric techniques, are needed.<sup>50</sup> In any event, continued sophisticated misuse of biased and irrelevant data is not the answer.

<sup>50</sup> Scherer's (1979b) delineation of the reasons for increased concentration both illustrates the type of analysis I have in mind and further supports the conclusion that statistical studies which associate variables aggregated by FTC-SIC-defined industries are fraught with error and a considerable potential for misinterpretation.

## REFERENCES

- Ayanian, Robert, "Advertising and Rate of Return," *Journal of Law & Economics*, October 1975, 18, 479-506.
- Bain, Joe S., "Relation of Profit Rate to Industry Concentration, American Manufacturing, 1930-1940," *Quarterly Journal of Economics*, August 1951, 65, 293-324.
- Benston, George J., "The Optimal Banking Structure: Theory and Evidence," *Journal of Bank Research*, Winter 1973, 3, 220-76.
- , "The Baffling New Numbers Game at the FTC," *Fortune*, October 1975, 92, 174-77.
- , "The Federal Trade Commission's Line of Business Report Program: A Benefit-Cost Analysis," in Harvey Goldschmid, ed., *Business Disclosure: Government's Need to Know*, New York: McGraw-Hill, 1979, 58-118.
- , "Accounting Numbers and Economic Values," *Antitrust Bulletin*, Spring 1982, 27, 161-215.
- , "The Costs of Complying with a Government Data Collection Program: The FTC's Line of Business Report," *Journal of Accounting and Public Policy*, Summer 1984, 3, 123-37.
- Bock, Betty, "Line of Business Reporting: A Quest for a Snark?," *The Conference Board Record*, November 1975, 12, 10-19.
- Breit, William and Elzinga, Kenneth G., "Information for Antitrust and Business Activity: Line-of-Business Reporting," in Kenneth W. Clarkson and Timothy J. Muris, eds., *The Federal Trade Commission Since 1970: Economic Regulation and Bureaucratic Behavior*, Cambridge: Cambridge University Press, 1981, 98-120.
- Buchanan, James M., *Cost and Choice*, Chicago: University of Chicago Press, 1969.
- and Thirlby, G. F., *L.S.E. Essays on Cost*, London: Weidenfeld and Nicolson, 1973.
- Comanor, William S. and Wilson, Thomas A., "The Effect of Advertising on Competition: A Survey," *Journal of Economic Literature*, June 1979, 17, 453-76.
- Demsetz, Harold, "Industry Structure, Market Rivalry, and Public Policy," *Journal of Law & Economics*, April 1973, 16, 1-9.
- , "Two Systems of Belief About Monopoly," in Harvey Goldschmid et al., eds., *Industrial Concentration: The New Learning*, Boston: Little, Brown and Co., 1974, 164-84.
- Fisher, Franklin M. and McGowan, John J., "On the Misuse of Accounting Rates of Return to Infer Monopoly Profits," *American Economic Review*, March 1983, 73, 82-97.
- Gale, Bradley T. and Branch, Ben S., "Concentration versus Market Share: Which Determines Performance and Why Does It Matter?," *Antitrust Bulletin*, Spring 1982, 27, 83-105.
- Hagerman, Robert and Zmijewski, Mark E., "Some Economic Determinants of Accounting Policy Choice," *Journal of Ac-*

- counting and Economics*, August 1979, 1, 141–61.
- Hall, Marshall and Weiss, Leonard, "Firm Size and Profitability," *Review of Economics and Statistics*, August 1967, 47, 319–31.
- Holthausen, Robert W. and Leftwich, Richard W., "The Economic Consequences of Accounting Choice: Implications of Costly Contracting and Monitoring," *Journal of Accounting and Economics*, August 1983, 5, 77–117.
- Liebowitz, Stanley J., "What Do Census Price-Cost Margins Measure?," *Journal of Law & Economics*, October 1982, 25, 231–46.
- \_\_\_\_\_, "On the Measurement of Monopoly Power," manuscript, University of Rochester, 1984.
- Long, William F., "Advertising Intensity, Market Share, Concentration and Degree of Cooperation," manuscript, Federal Trade Commission, September 1980.
- \_\_\_\_\_, (1981a) "Impact of Alternative Allocation Procedures on Econometric Studies of Structure and Performance," manuscript, Federal Trade Commission, July 1981.
- \_\_\_\_\_, (1981b) "The Effects of Inter-Firm Cooperation and Economies of Scale on Product Improving Research and Development Expenditures," manuscript, Federal Trade Commission, December 1981.
- \_\_\_\_\_, et al., *Volume I: Staff Analysis of the Benefits and Costs of the Federal Trade Commission's Line of Business Report*, Washington: Federal Trade Commission, September 1982.
- McGee, John, *In Defense of Industrial Concentration*, New York: Praeger, 1971.
- Martin, Stephen, (1981a) "Market, Firm, and Economic Performance: An Empirical Analysis," (FTC, 1981), in *Monograph Series in Finance and Economics*, Monograph 1983–1, Salomon Brothers Center for the Study of Financial Institutions, Graduate School of Business Administration, New York University, 1983.
- \_\_\_\_\_, (1981b) "Modeling Profitability at the Line of Business Level," manuscript, Federal Trade Commission, August 1981.
- Mautz, Robert and Skousen, Fred, "Common Cost Allocation in Diversified Companies," *Financial Executive*, June 1968, 36, 15–17, 19–25.
- Mueller, Dennis C., "Economies of Scale, Concentration, and Collusion," manuscript, Federal Trade Commission, September 1980.
- Ornstein, Stanley I., "Empirical Uses of the Price-Cost Margin," *Journal of Industrial Economics*, December 1975, 24, 105–17.
- Peltzman, Sam, "The Gains and Losses From Industrial Concentration," *Journal of Law & Economics*, October 1977, 20, 229–63.
- Ravenscraft, David J., "Intracompany Transfer Pricing and Profitability," manuscript, Federal Trade Commission, December 1981.
- \_\_\_\_\_, "Structure-Profits Relationships at the Line of Business and Industry Levels," (FTC, March 1982), *Review of Economics and Statistics*, February 1983, 65, 22–31.
- Salamon, Gerald L., "Models of the Relationship Between the Accounting and Internal Rate of Return: An Examination of the Methodology," *Journal of Accounting Research*, Spring 1973, 11, 296–303.
- \_\_\_\_\_, "Accounting Rate of Return, Measurement Error, and Tests of Economic Hypotheses: The Case of Firm Size," *American Economic Review*, June 1985, forthcoming.
- Scherer, F. M., (1979a) "Segmental Financial Reporting: Needs and Trade-Offs," in Harvey Goldschmid, ed., *Business Disclosure: Government's Need to Know*, New York: McGraw-Hill, 1979, 3–57.
- \_\_\_\_\_, (1979b) "The Causes and Consequences of Rising Industrial Concentration," *Journal of Law & Economics*, April 1979, 22, 191–211.
- \_\_\_\_\_, *Industrial Market Structure and Market Performance*, Boston: Houghton Mifflin Co., 1980.
- \_\_\_\_\_, "The Propensity to Patent," (FTC, September 1980), *International Journal of Industrial Organization*, November 1983, 1, 107–28.
- \_\_\_\_\_, (1981a) "Using Linked Patent and R&D Data to Measure Inter-Industry Technology Flows," (FTC, September 1981), in Zvi Griliches, ed., *R&D, Patents and Productivity*, NBER Conference Report, Chicago: University Chicago Press, 1984, 417–61.

- \_\_\_\_\_, (1981b) "Inter-Industry Technology Flows and Productivity Growth," (FTC, September 1981), *Review of Economics and Statistics*, November 1982, 64, 627-34.
- Schwert, G. William, "Using Financial Data to Measure Effects of Regulation," *Journal of Law and Economics*, April 1981, 24, 121-58.
- Scott, John T., "Multimarket Contact and Economic Performance," (FTC, April 1981), *Review of Economics and Statistics*, August 1982, 64, 368-75.
- Solomon, Ezra, "Alternative Rate of Return Concepts and Their Implications for Utility Regulation," *Bell Journal of Economics*, Spring 1970, 1, 65-81.
- Stauffer, Thomas A., "The Measurement of Corporate Rates of Return: A Generalized Formulation," *Bell Journal of Economics*, Autumn 1971, 2, 434-69.
- Watts, Ross and Zimmerman, Jerold, "Towards a Positive Theory of the Determination of Accounting Standards," *Accounting Review*, January 1978, 53, 112-34.
- Weiss, Leonard W., "Advertising, Profits, and Corporate Taxes," *Review of Economics and Statistics*, November 1969, 51, 421-30.
- \_\_\_\_\_, "Statement," in *Proceeding of the Conference on Bank Structure and Competition*, Federal Reserve Bank of Chicago, December 1970, 71-74.
- \_\_\_\_\_, "The Concentration-Profits Relationship in Antitrust," in Harvey Goldschmid et al., eds., *Industrial Concentration: The New Learning*, Boston: Little, Brown and Co., 1974, 184-233.
- \_\_\_\_\_, "The Structure-Conduct-Performance Paradigm and Antitrust," *University of Pennsylvania Law Review*, April 1979, 127, 1104-40.
- \_\_\_\_\_, and Pascoe, George, "Some Early Results on the Concentration-Profits Relationship from the FTC's Line of Business Data," manuscript, Federal Trade Commission, September 1981.
- \_\_\_\_\_, Martin, Stephen and Pascoe, George, "The Size of Selling Costs," (FTC, September 1981), *Review of Economics and Statistics*, November 1983, 65, 668-71.
- Williamson, Oliver E., "Economies as an Antitrust Defense: The Welfare Trade-Offs," *American Economic Review*, March 1968, 58, 18-36.
- Federal Trade Commission, Bureau of Economics Staff Memorandum, 1974 Form LB Revisions, Washington 1975.
- \_\_\_\_\_, (1981a) *Statistical Report: Annual Line of Business Report*, 1974, Washington 1981.
- \_\_\_\_\_, (1981b) *Statistical Report: Annual Line of Business Report*, 1975, Washington 1981.
- \_\_\_\_\_, *Statistical Report: Annual Line of Business Report*, 1976, Washington 1982.

# Do Large Deficits Produce High Interest Rates?

By PAUL EVANS\*

There has been much concern recently that large U.S. deficits have, or will soon, produce high interest rates, thus hindering capital formation and economic growth in the United States and the rest of the world. This view seems to be held not only by the popular press, but also by many respected economists (for example, Martin Feldstein, 1983).

In this paper, I survey U.S. history to determine whether this concern has merit. There are three periods during which the federal deficit has exceeded 10 percent of national income. In none of these periods did interest rates rise appreciably. Regression analysis applied to data from these three periods has not uncovered a positive association between deficits and interest rates. There also appears to be no evidence for a positive association between deficits and interest rates during the postwar period. I conclude from this survey that the concerns of the popular press and many economists may be misplaced.

## I. The Conventional Macroeconomic Paradigm

Perhaps the most widely used paradigm in macroeconomics is the *IS-LM* model. Its widespread use reflects not merely that it is analytically tractable, but also that many economists generally agree with its structural

assumptions and with its implications, one of which is that increased deficits raise interest rates. In this section, I briefly point out the implications of the *IS-LM* model. I then discuss some of the problems that arise in estimating how strong the relationship is between deficits and interest rates and in testing whether increased deficits do indeed raise interest rates.

Because the *IS-LM* model is so familiar, I need not derive its implications here. I simply assert that it implies a relationship between "the" nominal interest rate  $R$  and real government spending  $G$ , the real deficit  $D$ , the real money stock  $M/P$ , the expected inflation rate  $\Pi$ , an error term  $US$  that measures autonomous private spending, and an error term  $UM$  that measures the level of autonomous money demand. To keep the analysis simple, it is assumed that the relationship is linear and purely contemporaneous:

$$(1) \quad R = a_0 + a_1G + a_2D + a_3(M/P) + a_4\Pi + a_5US + a_6UM,$$

where the  $a$ 's are parameters. Of course, in an empirical analysis of real-world data, dynamics would be important so one would have to treat the  $a$ 's here and the  $b$ 's and  $c$ 's below as polynomials in the lag operator. Therefore, what I say below about the  $a$ 's,  $b$ 's, and  $c$ 's actually applies to the long-run multipliers of  $G$ ,  $D$ , and  $M/P$ .

According to the paradigm, at any given nominal interest rate, total spending increases (i.e., the *IS* curve shifts rightward) when government spending rises while the deficit is kept constant (i.e., government spending and taxes rise by the same amount); when the deficit rises while government spending remains constant (i.e., taxes fall); when the expected inflation rate rises (i.e., the real interest rate falls); or when autonomous spending rises. Therefore, since the

\*Department of Economics, University of Houston, Houston, TX, 77004. I am grateful to Moses Abramovitz, Armen Alchian, Bob Barro, Martin Bailey, Michael Boskin, Charles Calomiris, Alex Field, Milton Friedman, Bob Hall, Doug Joines, Bob Lucas, Tom Mayer, Ben McCallum, John Seater, two referees, and the participants in seminars at Claremont College, the Federal Reserve Bank of San Francisco, Stanford University, Boston College, Ohio State University, Michigan State University, University of California-Santa Cruz, Swarthmore College, University of California-Irvine, the Federal Reserve Bank of St. Louis, Washington University, University of Chicago, and Lehrman Institute.



*LM* curve is upward sloping,  $a_1, a_2, a_4$ , and  $a_5$  are positive. The parameter  $a_3$  is negative and the parameter  $a_6$  is positive because decreasing the real money stock or increasing the autonomous demand for money lowers the spending that can be financed at any given nominal interest rate (i.e., shifts the *LM* curve leftward), thereby driving up the nominal interest rate.

Because the expected inflation rate  $\Pi$  is unobservable but may be systematically related to other variables, one should eliminate it from equation (1). Suppose that it is related to government spending, the deficit and the real money stock by

$$(2) \quad \Pi = b_0 + b_1 G + b_2 D + b_3 (M/P) + U\Pi$$

where the  $b$ 's are parameters and  $U\Pi$  is  $\Pi$ 's unsystematic part. Substituting equation (2) into equation (1) then yields

$$(3) \quad R = c_0 + c_1 G + c_2 D + c_3 (M/P) + U,$$

where

$$(4) \quad c_0 = a_0 + a_4 b_0,$$

$$(5) \quad c_1 = a_1 + a_4 b_1,$$

$$(6) \quad c_2 = a_2 + a_4 b_2,$$

$$(7) \quad c_3 = a_3 + a_4 b_3,$$

$$(8) \quad U = a_4 U\Pi + a_5 US + a_6 UM,$$

$$a_4, a_5, a_6 > 0.$$

Suppose that one estimates equation (3) consistently, treating  $U$  as an unobservable and unsystematic error term. The estimates converge in probability to  $a_1 + a_4 b_1$ ,  $a_2 + a_4 b_2$ , and  $a_3 + a_4 b_3$ , rather than  $a_1$ ,  $a_2$ , and  $a_3$ . For many purposes, this causes no trouble. For example, if one wants to know what an increased deficit does to nominal interest rates,  $a_2 + a_4 b_2$  is the right answer. On the other hand, if one wants to know what a larger deficit does to real interest rates,  $a_2$  is the right answer. Nevertheless, if  $b_1$ ,  $b_2$ , and  $b_3$  are nonnegative, one's estimates of  $c_1$ ,  $c_2$ , and  $c_3$  provide upper bounds for  $a_1$ ,  $a_2$ , and

$a_3$ . Therefore, if  $a_1 > 0$  and  $a_2 > 0$ , one should find that  $c_1 > 0$  and  $c_2 > 0$ . It is natural to assume that the  $b$ 's are nonnegative because in the *IS-LM* model increasing  $G$ ,  $D$ , or  $M/P$  raises aggregate demand and hence future inflation if the price level evinces any stickiness. Moreover, increased  $G$  or  $D$  may force increased money growth in the future, thereby raising the inflation rate. In such circumstances, reasonable behavior in forming expectations rules out negative  $b$ 's.

Even if government spending, the deficit, and the nominal money stock are completely exogenous variables, the ordinary least squares estimates  $\hat{c}_1$ ,  $\hat{c}_2$ , and  $\hat{c}_3$ , may not be consistent. The reason is that the error terms  $U\Pi$ ,  $US$ , and  $UM$  may have nonnegligible contemporaneous effects on the price level. For example, increasing  $U\Pi$  or  $US$  raises aggregate demand, simultaneously raising the nominal interest rate and putting upward pressure on the price level. If the price level should actually rise before  $U\Pi$ 's or  $US$ 's influence on the nominal interest rate has dissipated,  $M/P$  falls endogenously, tending to bias  $\hat{c}_3$  downward asymptotically. In contrast, increasing  $UM$  raises the interest rate but lowers aggregate demand, putting downward pressure on the price level. Consequently,  $\hat{c}_3$  may in fact be biased upward.<sup>1</sup> This bias is more serious, the less sticky the price level.

Ordinary least squares estimates of  $c_1$ ,  $c_2$ , and  $c_3$  may also be inconsistent because government spending, the deficit, and the nominal money stock may be correlated with the error term  $U$ . To see how this can happen, consider first increasing  $U\Pi$  or  $US$ . The nominal interest rate and output rise simultaneously. The increased output reduces some components of government spending, increases tax revenue and thus shrinks the deficit. Furthermore, the monetary authorities may accommodate some of the increased money demand that the higher spending induces. As a result,  $R$  rises while  $G$  and  $D$  are falling and  $M$  is rising endogenously.

<sup>1</sup>For simplicity, I assume that  $G$ ,  $D$ , and  $M$  are orthogonal so that one can treat each coefficient's bias in isolation.

Therefore, variance in  $UII$  or  $US$  biases  $\hat{c}_1$  and  $\hat{c}_2$  downward while biasing  $\hat{c}_3$  upward. Next, consider increasing  $UM$ . As the nominal interest rate is rising, output falls, leading government spending and the deficit to rise. The monetary authorities probably accommodate some of the extra autonomous money demand; hence  $M/P$  rises. Consequently, variance in  $UM$  biases  $\hat{c}_1$ ,  $\hat{c}_2$ , and  $\hat{c}_3$  upward. All in all, the endogeneity of  $G$ ,  $D$ ,  $M$ , and  $P$  spells inconsistent least squares estimates of  $c_1$ ,  $c_2$ , and  $c_3$ . The inconsistencies can be serious and of either sign, depending on how important is each source of endogeneity.

One must somehow overcome the problem of inconsistency detailed above. In this paper, I have overcome it in three ways. First, in three of my four sample periods, exogenous influences dominate the sample variances of government spending, tax revenue, and the nominal money stock. Second, I have used monthly data wherever possible so that I can reasonably take the real money stock as exogenous if the nominal stock is. Third, if I cannot reasonably take government spending, tax revenues, and the nominal or real money stock as exogenous, I estimate using two-stage least squares.

## II. The U.S. Experience during the Civil War

The Civil War broke out in 1861 and ended in 1865. Federal spending rose from 1.6 percent of national income in the four fiscal years prior to the war to 15.5 percent in the fiscal years 1862–65.<sup>2</sup> Little of the federal war effort was financed by taxes. Consequently, the deficit, which had averaged 0.5 percent of national income in the four fiscal years prior to the war, jumped to 12.1 percent during the war.<sup>3</sup> About one-

quarter of the deficit was monetized in 1862 and 1864, and over one-half was monetized in 1863. By 1865, however, the Union had succeeded in reducing monetization to 10 percent even though the deficit was appreciably larger than it had been in the previous three fiscal years.

Anyone taking the conventional paradigm seriously would predict from the above facts that interest rates must have risen sharply during the Civil War, especially after monetization fell in fiscal year 1865. Monetization might have held interest rates down somewhat, especially in fiscal year 1863. Because, however, the price level rose a great deal during the war,<sup>4</sup> it is unlikely that interest rates were held down much.

Figure 1 shows that this prediction is wide of the mark. I have plotted the deficit as a percent of national income and two interest rates, measured in percentage points per annum. The deficit ratio ( $DR$ ) jumped sharply during the war. Neither the commercial paper rate ( $CPR$ ), nor the railroad bond rate ( $RRBR$ ), shows any obvious correlation with the deficit ratio.<sup>5</sup> Indeed, the railroad bond rate tended to fall during most of the war. The commercial paper rate jumped in November 1860 upon President Lincoln's election, but fell as soon as the public had satisfied its increased liquidity preference.<sup>6</sup> It

---

real capital loss that a rising price level imposes on the holders of federal debt, add the real capital gains that indexing bestows, and add any rise in the market value of the debt. I use the official measure because it tells essentially the same story as a corrected measure would.

The first two corrections swell the *ex post* Civil War deficits by an average of 1.0 percent of national income. The reason is that much of the interest-bearing federal debt was indexed to gold (Wesley Mitchell, 1903) and the greenback price of gold rose more than the price level did. Because, however, no one at the time could be certain that interest and principal would actually be paid in gold, this correction is probably too large for the *ex ante* deficit.

<sup>4</sup>The Hoover *CPI*, which had fallen 0.7 percent a year between 1858 and 1861, rose 6 percent in 1862, 16 percent in 1863, 22 percent in 1864, and 11 percent in 1865 (all being fiscal years).

<sup>5</sup>The same is true of the New England municipal bond rate, an alternative measure of the long-term interest rate.

<sup>6</sup>Fear of war is well known to raise the demand for money.

<sup>2</sup>The years referred to in the paper are all fiscal years that run from July 1 to June 30. A data appendix, available from the author upon request, describes the data cited here, provides their sources and lists them.

<sup>3</sup>Here and elsewhere in the text, I have used the official measure of the federal deficit. It is well known, however, that this measure is incomplete when the price level is changing, when the federal debt is indexed, or when its market value changes. To obtain the correct measure, one must subtract from the official measure the

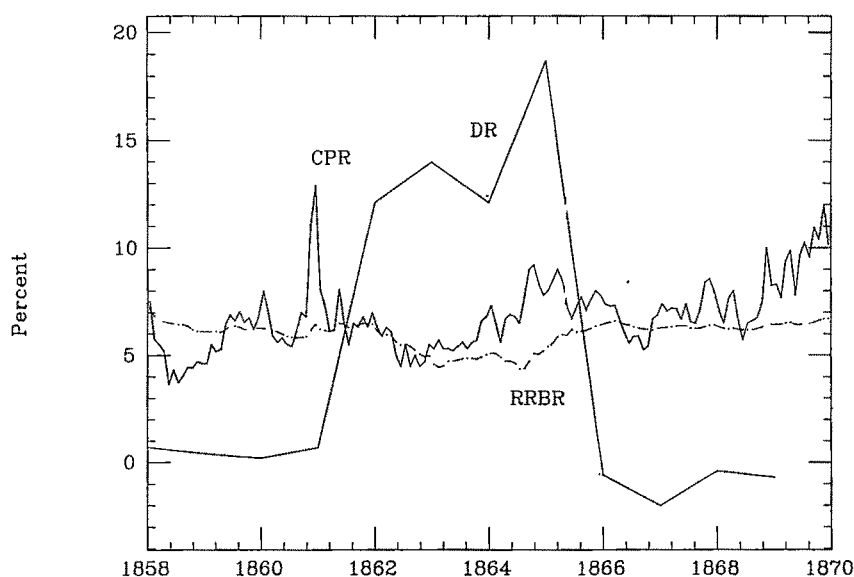


FIGURE 1. U.S. EXPERIENCE DURING THE CIVIL WAR

also drifted upward as the end of the war approached, perhaps reflecting the liquidity demand of southerners in anticipation of a Union victory and Reconstruction. In any case, the movement of the commercial paper rate was not closely related to the deficit.

Although the large Civil War deficits were not associated with sharply higher interest rates, they may nevertheless have had some effect. To determine how the deficits affected interest rates, I used two-stage least squares to fit the following regressions to annual data spanning the fiscal-year period 1858-69:<sup>7</sup>

$$(9) \quad \hat{CPR}_t = 7.48 + .583GR_t - .440DR_t$$

(1.05) (.364)      (.245)

$$- .373MR_t$$

(.291)

$$S.E. = 1.02, D-W = 2.29;$$

$$(10) \quad \hat{RRBR}_t = 7.42 + .295GR_t - .242DR_t$$

(0.26) (.090)      (.061)

$$- .300MR_t$$

(.072)

$$S.E. = 0.25, D-W = 1.75;$$

$$(11) \quad \hat{NEBR}_t = 5.52 + .342GR_t - .262DR_t$$

(0.16) (.055)      (.037)

$$- .222MR_t$$

(.044)

$$S.E. = 0.15, D-W = 2.17;$$

<sup>7</sup>I treated  $MR$  as endogenous, using as its instrument the nominal stock of paper money deflated by the Consumer Price Index lagged one year and trend real national income. Ordinary least squares, however, yielded similar results.

where  $t$  is an index of time;  $CPR$ ,  $RRBR$ , and  $NEBR$  are, respectively, the commercial paper rate, railroad bond rate, and New England municipal bond rate;  $GR$  is the ratio of real federal spending to trend real national income;  $DR$  is the ratio of the real deficit to trend real national income;  $MR$  is the ratio of the real stock of paper money outstanding at the end of each fiscal year to trend real national income; and standard

errors are shown in parentheses below each estimated coefficient.<sup>8</sup>

If one can interpret equations (9)–(11) as estimates of equation (3), *GR* and *DR* should have positive coefficients.<sup>9</sup> The ratio *GR* does indeed have positive coefficients, and two of the three are statistically significant at any reasonable level. In contrast, *DR* has negative coefficients with *t*-statistics  $-1.79$ ,  $-3.98$ , and  $-7.07$ . Since these *t*-statistics are significant at the 10, 1, and .1 percent levels, they provide strong evidence against the conventional paradigm. They also show that large Civil War deficits did not produce high interest rates. If anything, interest rates would have been lower, had the Union levied lower taxes, yielding a larger deficit.

The evidence is clearly less sophisticated than one would like to use. In particular, the regressions are based on only twelve observations and fairly crude data. Nevertheless, the evidence is fairly strong because the enormous sample variance of the independent variables reduces the need for a large sample.

The conventional paradigm has apparently failed. One therefore needs an explanation of its failure in order either to justify its continued use or to provide an alternative to it. I can think of six potential explanations for why federal deficits did not raise interest rates during the Civil War. First, the inflation premia in the nominal interest rates *CPR*, *RRBR*, and *NEBR* may have fallen as the deficit ratio rose. In terms of the analysis of Section I,  $b_2$  may have been negative enough to make  $c_2$  negative. A larger deficit may have lowered inflation expectations because the public may have believed Congress and the Lincoln Administration when they promised to return to the gold standard at

the prewar parity. After the gold standard was abandoned in December 1861, each increase in the deficit and hence in the price level meant that the price level had to fall by more to reattain the prewar parity.

This explanation makes sense for long-term interest rates like the railroad and New England municipal bond rates, but makes much less sense for short-term interest rates like the commercial paper rate. Probably households did not think that deflation would commence, or even that inflation would slow down much while the war was still going on. More likely, they expected the inflation rate to remain fairly close to what it was at the time.<sup>10</sup> If so, real *ex ante* short-term interest rates can be adequately proxied by the *ex post* real yields. These yields should then be increasing functions of the deficit ratio.<sup>11</sup>

<sup>10</sup>Using yields on greenback and gold government bonds, Richard Roll (1972) has calculated a one-year-ahead expected growth rate of the greenback price of gold for each week between May 2, 1863, and December 12, 1864. His series implies that even in 1863, when no one could reasonably have thought that the Civil War would soon end, the public expected the price of gold to fall over the next year, often by appreciable amounts. Although the price of gold is not entirely representative of prices in general, this evidence does suggest that the public actually expected deflation and that the argument of the previous paragraph might hold even for short-term interest rates. Because Roll based his estimates on strong assumptions, however, the hypothesis here still seems like a reasonable one.

<sup>11</sup>Given the hypothesis that actual and expected inflation rates differed by an independently and identically distributed error term, the estimated coefficients on *GR*, *DR*, and *MR* converge in probability to  $a_1 - (1 - a_4)b_1$ ,  $a_2 - (1 - a_4)b_2$ , and  $a_3 - (1 - a_4)b_3$ . In the standard *IS-LM* model, the parameter  $a_4$  is approximately  $1/[(1 + \beta/\alpha)(1 - \tau)]$ , where  $\alpha$ ,  $\beta$ , and  $\tau$  are the slope of the *LM* curve, minus the slope of the *IS* curve and the effective marginal tax rate on nominal interest income, respectively. Since  $\tau$  was close to zero during the Civil War,  $a_4$  lies between 0 and 1. If  $b_2$  is negative as the previous paragraph has argued it might be, the coefficient on *DR* is an upper bound to  $a_2$ . In this case, rejecting the hypothesis that this coefficient is positive is equivalent to rejecting the hypothesis that  $a_2$  is. In the more likely case that  $b_2$  is positive, however, failure to reject the null hypothesis is not evidence against the hypothesis that  $a_2 > 0$ . In any case, a positive  $a_2$  implies either that  $c_2$  is positive or that the coefficient on *DR* in the regression for the real interest rate is positive. One can obtain similar restrictions on the estimated coefficients of *GR* and *MR*.

<sup>8</sup>Throughout the empirical analysis, I have measured interest rates and ratios in percentage points. For example, equation (11) implies that increasing government spending by 1 percent of national income raised the New England municipal bond rate by .342 percentage points.

<sup>9</sup>Actually, the equation has *GR*, *DR*, and *MR* in it, rather than *G*, *D*, and *M/P*. In choosing this functional form, I am imposing the restriction of homogeneity; i.e., that doubling *G*, *D*, *M/P*, and the scale of the U.S. economy does not affect the equilibrium nominal and real interest rates.

To test this hypothesis, I fitted the following equation:<sup>12</sup>

$$(12) \quad RC\hat{P}R_t = 19.22 + 4.62GR_t - 4.15DR_t - 3.80MR_t, \\ (4.32) \quad (1.50) \quad (1.01) \quad (1.20)$$

$$S.E. = 4.22, D-W = 2.81;$$

where  $RCPR$  is  $CPR$  less the rate of consumer price inflation. Note that the coefficients on  $GR$  and  $MR$  support the conventional paradigm but the coefficient on  $DR$  does not. Moreover,  $DR$ 's coefficient is statistically significant at the .01 level. These results are similar to those for the nominal interest rates.

Second, usury ceilings (typically 6 percent; see Sidney Homer, 1963) may have prevented private borrowers from competing with the federal government. Again, it is hard to take this argument seriously, given the many ways borrowers and lenders have always found to get around usury laws. In particular, selling securities at discount was legal.

Third, patriotism may have spurred households to save enough more to finance the deficits. In empirical economics, however, appealing to variables as difficult to measure as patriotism is the last refuge of a scoundrel. Nevertheless, in Section IV, I do try to test this hypothesis, using a proxy for patriotism. I find no evidence that patriotism spurs saving.

Fourth, capital inflows from abroad may have financed the deficit.<sup>13</sup> Capital inflows, however, never exceeded 10 percent of the deficit during the Civil War.

Fifth, wartime disruptions may have reduced the profitability of investment, thereby reducing aggregate demand and interest rates.<sup>14</sup> Stanley Engerman (1966) has shown

that these disruptions were important and that they may have depressed investment appreciably. These facts, however, explain why interest rates were not higher on average during the Civil War, but not why interest rates were negatively related to the deficits. Many of the disruptions began at the start of the war and remained more or less constant until sometime after the war was over. For example, textile manufacturers lost their southern markets when the war started and did not regain them until well after Reconstruction had begun. Since the effect of the disruptions on aggregate demand was probably not very correlated with the deficit, they provide an unsatisfactory explanation for  $DR$ 's negative coefficients.

Sixth, private saving may automatically rise to finance increased deficits when the government is perceived to service the debt with future taxes. I elaborate further on this explanation in Sections IV and VI.

I have estimated analogues to regressions (9)–(12) in which  $DR_t + MF_{t-1}(1/P_t - 1/P_{t-1}) + IBDEBT_{t-1}(PG_t/P_t - PG_{t-1}/P_{t-1})$  replaced  $DR_t$ , where  $MF$  is the stock of federal paper money at the end of the fiscal year;  $IBDEBT$  is the interest-bearing federal debt at the end of the fiscal year, all of which I assume was indexed to gold;  $P$  is the average Consumer Price Index in the fiscal year; and  $PG$  is the average greenback price of gold in the fiscal year. The terms added to  $DR$  seem to serve largely as measurement error, yielding coefficients that are smaller in magnitude and less significant than those reported for  $DR$  above but that have the same sign. Therefore, "correcting" the official measure of the deficit does not change the conclusions reached above. Nor are they changed by replacing  $GR$ ,  $DR$ , and  $MR$  with  $\ln G$ , the logarithm of real tax revenue and  $\ln(M/P)$ .

### III. The U.S. Experience during World War I

In April 1917, the United States entered World War I, which lasted until November 1918. Federal spending rose rapidly from an average of 1.8 percent of  $GNP$  in (fiscal years) 1913–16 and the first three quarters of 1917 to 9.0 percent in the last quarter of

<sup>12</sup> The Durbin-Watson statistic does not imply statistically significant serial correlation since there are only eight degrees of freedom.

<sup>13</sup> Milton Friedman has suggested this possibility to me.

<sup>14</sup> I owe this point to Alex Field.

1917, to 18.5 percent in 1918 and to 23.0 percent in 1919.

The federal government financed more of this war with taxes than it had the Civil War. Nevertheless, the deficit rose from an average of 0.0 percent of *GNP* in the four fiscal years prior to the war to 6.4 percent in 1917:4, 13.2 percent in 1918, and 16.6 percent in 1919.<sup>15</sup> The Federal Reserve monetized 26.2 percent of the deficit in 1917:4, but reduced that figure to 8.9 percent in 1918 and to only 4.7 percent in 1919. Thus, the deficit was much less monetized during World War I than during the Civil War.

According to the conventional paradigm, the large deficits of World War I should have driven interest rates up sharply, especially in 1918 and 1919 when there was little monetization. Figure 2 shows that this prediction is just as wide of the mark for World War I as it was for the Civil War. As in Figure 1, I have plotted the commercial paper rate, the railroad bond rate, and the deficit ratio, labeling them *CPR*, *RRBR*, and *DR*, respectively. It is hard to see any correlation between *DR* and either *CPR* or *RRBR*. The railroad bond rate was rather stable.<sup>16</sup> The commercial paper rate moved around somewhat, but these movements were largely unrelated to the deficit ratio. In particular, the bulge in 1913 reflected a panic nipped in the bud by the issue of Aldrich-Vreeland currency;<sup>17</sup> and the bulge in 1914 followed the outbreak of World War I in Europe.

To show more rigorously that large deficits did not produce high interest rates during World War I, I have fitted the commercial paper rate and the railroad bond rate to constant terms, time trends, six lagged values of the respective dependent variable, and the current and six lagged values of *GR*, the ratio of real federal spending to trend real

*GNP*; *DR*, the ratio of the real deficit to trend real *GNP*; and *MR*, the ratio of the real *M2* money stock to trend real *GNP*.<sup>18</sup> I applied ordinary least squares to monthly observations spanning the period January 1914 to December 1920. Tables 1 and 2 report these regressions.

Although the *F*-statistics in Table 1 do not reveal a statistically significant relationship between the commercial paper rate and the right-hand side variables, there is one nonetheless. The coefficients for *GR* and *DR* sum to .261 and -.320, and have *t*-statistics 2.14 and -2.25. These sums are thus statistically significant at the .05 level. The conventional paradigm is consistent with the observed positive sum for *GR*'s coefficients, but is inconsistent with the observed negative sum for *DR*'s coefficients.

The *F*-statistics for *GR* and *DR* in Table 2 imply highly significant relationships. Furthermore, the coefficients for *GR* and *DR* sum to .0582 and -.0684, and have *t*-statistics 2.29 and -2.32. Again, the sums are statistically significant, the positive sum for *GR* is consistent with the conventional paradigm, and the negative sum for *DR* is not.

Table 3 provides further evidence that the conventional paradigm fails. The table reports the regression that I have fitted to *RCPR*, the *ex post* real commercial paper rate.<sup>19</sup> I assume that the estimated coefficients in the table can tell one something

<sup>15</sup>Adjusting the fiscal year 1918 and 1919 figures for inflation as discussed in fn. 4 reduced them by about 1.3 percent of *GNP*.

<sup>16</sup>The period August 1914 to November 1914 might have seen some movement in this interest rate; one cannot know because the bond markets were officially closed.

<sup>17</sup>See Milton Friedman and Anna Schwartz (1963, p. 172).

<sup>18</sup>See the data appendix for the definitions of these variables. Note that these regressions are monthly analogues to those estimated for the Civil War. I have included the lagged variables in order to capture the lag structure that an *IS-LM* model more sophisticated than the one analyzed in Section I would imply. The lag length 6 seems to estimate the lag structures well. The results reported in the text are hardly changed, however, by using 3, 9, and 12 lags, or by using polynomial distributed lags. The results are also little changed if one instruments for *MR*, using *M2* deflated by the previous month's wholesale price index and trend real *GNP*.

<sup>19</sup>I have assumed that the Federal Reserve obtained its series on the 4-6-month commercial paper rate by averaging the rates on paper divided equally among the 4, 5, and 6-month maturities. I have therefore used the formula  $RCPR_t = CPR_t - (1200/3)[\ln(P_{t+6}/P_t)/6 + \ln(P_{t+5}/P_t)/5 + \ln(P_{t+4}/P_t)/4]$ , where *P* is the wholesale price index.

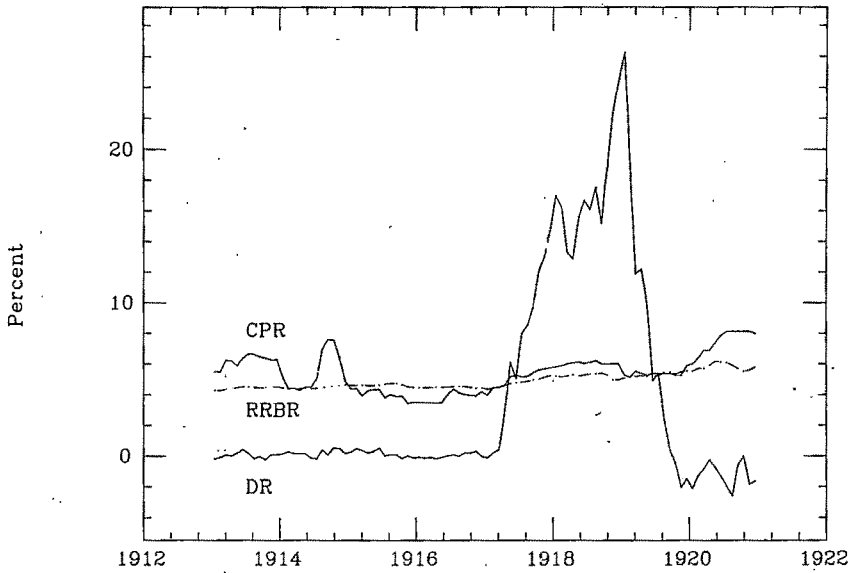


FIGURE 2. U.S. EXPERIENCE DURING WORLD WAR I

TABLE 1—THE CPR REGRESSION FOR WORLD WAR I

Lag	Right-Hand Side Variables <sup>a</sup>			
	CPR	GR	DR	MR
0		.029 (.100)	.008 (.103)	.057 (.088)
1	1.225 (.136)	.124 (.099)	-.155 (.116)	-.191 (.149)
2	-.479 (.215)	.129 (.108)	-.097 (.118)	.126 (.154)
3	.071 (.216)	.112 (.108)	-.113 (.119)	.156 (.161)
4	-.209 (.211)	-.004 (.096)	.010 (.115)	-.301 (.164)
5	.205 (.206)	-.022 (.096)	-.017 (.117)	.146 (.182)
6	-.103 (.133)	-.106 (.089)	.054 (.095)	-.057 (.114)
Sum	.711 (.099)	.261 (.122)	-.320 (.142)	-.064 (.042)
F-Statistic	32.2	0.9	0.9	1.1
Marginal Significance Level	.00	.53	.49	.41
Constant	3.40 (1.78)	Time Trend	-.00882 (.00752)	
R <sup>2</sup> = .942	S. E. = 0.39		Q(27) = 19.0 <sup>b</sup>	P = .87

<sup>a</sup>Standard errors are shown in parentheses.

<sup>b</sup>The Box-Pierce statistic tests for serial correlated error terms. *P* is its marginal significance level.



TABLE 2—THE *RRBR* REGRESSION FOR WORLD WAR I

Lag	Right-Hand Side Variables <sup>a</sup>			
	<i>RRBR</i>	<i>GR</i>	<i>DR</i>	<i>MR</i>
0		.0351 (.0210)	-.0312 (.0211)	-.0148 (.0181)
1	1.303 (.128)	.0366 (.0192)	-.0268 (.0218)	.0418 (.0307)
2	-.516 (.216)	-.0306 (.0219)	.0483 (.0234)	-.0086 (.0307)
3	.264 (.210)	.0264 (.0224)	-.0626 (.0243)	-.0337 (.0308)
4	-.510 (.188)	.0383 (.0204)	-.0209 (.0252)	.0106 (.0316)
5	.440 (.198)	-.0383 (.0201)	-.0129 (.0231)	.0202 (.0333)
6	-.234 (.147)	-.0392 (.0199)	.0375 (.0199)	-.0075 (.0202)
Sum	.748 (.079)	.0582 (.0254)	-.0684 (.0294)	.0079 (.0078)
F-Statistic	52.3	2.7	3.9	0.7
Marginal Significance Level	.00	.02	.00	.68
Constant	0.70 (0.37)	Time Trend		.00282 (.00142)
$R^2 = .985$		$S.E. = 0.07$	$Q(27) = 13.0^b$	$P = .99$

<sup>a,b</sup>See Table 1.TABLE 3—THE *RCPR* REGRESSION FOR WORLD WAR I

Lag	Right-Hand Side Variables <sup>a</sup>			
	<i>RCPR</i>	<i>GR</i>	<i>DR</i>	<i>MR</i>
0		2.21 (0.93)	-2.28 (0.97)	-4.16 (1.23)
1	1.743 (.132)	2.11 (0.97)	-1.14 (1.17)	7.36 (1.51)
2	-.991 (.268)	-0.45 (1.10)	-0.58 (1.17)	-3.69 (1.83)
3	.506 (.318)	0.63 (1.03)	-0.12 (1.15)	3.31 (2.06)
4	-.734 (.339)	0.62 (0.96)	-1.46 (1.15)	-6.31 (2.42)
5	.419 (.318)	-1.78 (0.96)	2.39 (1.18)	2.48 (2.54)
6	-.051 (.210)	-0.30 (0.91)	-0.65 (0.95)	0.41 (1.84)
Sum	.891 (.109)	3.03 (1.01)	-3.85 (1.22)	-0.80 (0.39)
F-Statistic	248	2.2	2.5	7.4
Marginal Significance Level	.00	.05	.02	.00
Constant	27.3 (14.5)	Time Trend		-.175 (.069)
$R^2 = .990$		$S.E. = 3.57$	$Q(27) = 22.8^b$	$P = .70$

<sup>a,b</sup>See Table 1.



TABLE 4—REGRESSIONS FOR CONSUMPTION EXPENDITURES, 1901–29<sup>a</sup>

Equation	Constant	YD	D	Y-G	GT	FDEBT	S.E.	D-W
(4.1)	-3517 (1908)	.998 (.035)					2513	0.94
(4.2)	-2637 (1267)	.979 (.023)	-.737 (.121)				1658	2.02
(4.3)	-2637 (1267)		.242 (.126)	.979 (.023)			1658	2.02
(4.4)	-1439 (2116)		-.307 (.763)	.955 (.041)	.619 (.851)		1742	1.90
(4.5)	-2118 (1227)			.969 (.022)	.282 (.141)		1671	1.96
(4.6)	-3370 (1794)		.273 (.132)	.999 (.041)		-.0407 (.0617)	1704	1.99
(4.7)	-2492 (1701)			.981 (.039)	.291 (.146)	-.0284 (.0611)	1714	1.94
(4.8)				.932 (.006)	.265 (.141)		1671	1.73

<sup>a</sup>I treated *YD*, *D*, and *Y-G* as endogenous and all other right-hand side variables as exogenous. In all regressions, I used as instruments the constant term, Robert Gordon's natural output series, its products with the growth rate of the nominal *M2* money stock and with the marginal tax rate on income, the size of the military, and real government purchases. When I included *GT* and/or *FDEBT*, I added them to the list of instruments.

useful about how the *ex ante* real commercial paper rate responded to the federal spending ratio, the deficit ratio and the money ratio. (Footnote 11 provides some justification for so assuming.) Given this assumption, one can infer that the real commercial paper rate rose with the federal spending ratio because *GR*'s coefficients are jointly significant, summing to a highly significant 3.03. Furthermore, *RCPR* fell with the deficit ratio because its coefficients are jointly significant and sum to  $-3.85$ , which is highly significant. Finally, *RCPR* fell with the money ratio since *MR*'s coefficients are jointly significant and sum to  $-.796$ , which is statistically significant at the .05 level. The first and third finding are consistent with the conventional paradigm, but the second is not.

The paradigm fails just as badly for regressions in which  $\ln G$ , the logarithm of real tax revenue and  $\ln(M/P)$  replaced *GR*, *DR*, and *MR* and in regressions in which I replaced *DR* with  $DR_t - FDEBT_{t-1}(1/P_{t-1} - 1/P_t)$ , where *FDEBT* is end-of-month gross federal debt and *P* is the wholesale price index. This measure of the deficit excludes the real capital losses that inflation imposed on the holders of federal debt.

For the Civil War, I could only speculate on why the conventional macroeconomic paradigm failed. For World War I, data are available for determining where that paradigm went wrong. According to the paradigm, increasing the deficit while keeping government spending constant raises disposable income, thereby stimulating consumption at any given nominal and real interest rates. The World War I deficits, however, did not stimulate consumption spending much if at all. Table 4 reports the evidence for this assertion.

I fitted the regressions in Table 4 to annual data that span the period 1901 to 1929, using two-stage least squares. Regression (4.1) is the simplest possible consumption expenditure function.<sup>20</sup> It relates real consumption expenditure to *YD*, current real disposable income; which I measure as *Y*, real net national product; minus *G*, real

<sup>20</sup>It would be better to fit a consumption function, but only data on consumption expenditure are readily available. Consumption expenditure includes spending on consumer durables and does not include the service flow from the current stock of consumer durables. See the data appendix (available upon request) for more information.

government purchases of all branches of government; plus  $D$ , the real deficit of all branches of government.<sup>21</sup> This simple and conventional consumption expenditure function implies that, *ceteris paribus*, each dollar of deficit raised consumption expenditure by nearly a dollar. Consequently, the deficits during World War I should have driven interest rates up considerably.

Regression (4.2) shows that the current deficit did not affect consumption expenditure just through its effect on current disposable income.<sup>22</sup> The deficit has a highly significant, negative coefficient. In addition, the Durbin-Watson statistic jumps to 2.02 from 0.94 for regression (4.1), thus suggesting that regression (4.1) was misspecified by omitting  $D$ . The negative coefficient implies that an increased deficit raised consumption expenditure by much less than the marginal propensity to purchase consumption goods from  $Y - G$ . Households must therefore have treated  $Y - G$ , social disposable income, differently from  $D$ , the government's contribution to private disposable income—raising consumption expenditure much more when the former increased a given amount than when the latter did. Regression (4.3) rewrites regression (4.2) in a way that makes this distinction clear. It shows that the marginal propensity to purchase consumption goods is .979 from social disposable income, but is only .242 from the deficit.

Equation (4.3) suggests that households may have tried to spend some of the World War I deficits. (The coefficient on  $D$  is statistically significant at the .05 level.) Doing so would have tended to raise interest rates. Why then did interest rates not rise with the federal deficit? One answer consistent with the facts is that households really do have a positive marginal propensity to spend deficits, but that the relatively high marginal corporate profits tax rates in 1917–19 depressed the demand for business investment

by more than the deficits raised the demand for consumption expenditure.<sup>23</sup>

Robert Barro has provided a different answer. According to Barro (1974), households understand that a current deficit entails additional future tax liabilities equal in present value to the current deficit. For this reason, they do not regard a deficit as contributing to their permanent private disposable incomes any more than it is a contribution to social disposable income. Consequently, households do not try to spend deficits. The variable  $D$  in regression (4.3) has a statistically significant coefficient because  $D$  is highly correlated with the transitory part of  $G$ .<sup>24</sup> It is optimal for households facing any given current social disposable income to consume more, the higher the transitory part of  $G$ .<sup>25</sup> Therefore, the consumption expenditure function should have a statistically insignificant coefficient on the deficit if one has also included a good proxy for transitory government purchases.

Regression (4.4) shows the result of adding  $GT$ , my measure of transitory government purchases, to regression (4.3). (See the data appendix to learn how I constructed it.) The estimated coefficient on  $D$  loses all statistical significance and indeed becomes negative. Dropping  $D$  yields regression (4.5). The estimated coefficient on  $GT$  is positive and statistically significant, thus supporting Barro's hypothesis. The coefficient implies that increasing government purchases by a dollar lowered consumption expenditure by 28.2 cents less when the increase was transitory than when it was permanent.

Barro's (1981) analysis also implies that the transitory part of federal spending should affect interest rates much more than the permanent part. I have tried to test this implica-

<sup>21</sup>I am here following the advice of Michael Boskin (1978) in using private disposable income rather than personal disposable income.

<sup>22</sup>Lagged variables contributed no statistically significant explanatory power to equations (4.2)–(4.8).

<sup>23</sup>The effective corporate profits tax rate averaged 0.3 percent between 1901 and 1916, 32.8 percent between 1917 and 1919, and 12.1 percent between 1920 and 1929. See John Seater (1982, p. 363).

<sup>24</sup>See Barro (1979) for a theoretical rationalization for this correlation as well as empirical evidence that it is strong.

<sup>25</sup>This proposition is simply a corollary of the permanent income hypothesis. See Barro (1981) for a discussion of it.

tion but have had little success. It has proven hard to construct good monthly proxies for the permanent and transitory parts of the federal spending ratio  $GR$ .<sup>26</sup>

Regressions (4.6) and (4.7) provide evidence that households do not spend more on consumption goods when  $FDEBT$ , the beginning-of-year real stock of federal debt, is larger. In both regressions,  $FDEBT$  enters with a negative coefficient, which is, however, not statistically significant at the .05 level.

In regression (4.5) the marginal propensity to purchase consumption goods from permanent social disposable income is .969, a figure that seems too high. Furthermore, the constant term is negative but not statistically significant at the .05 level. I have therefore reestimated (4.5) excluding the constant term.<sup>27</sup> Regression (4.8) is the result. The marginal propensity is now the more reasonable value .932, and the coefficient on  $GT$  remains positive and statistically significant at the .05 level on a one-tailed test.

Barro has provided a consistent explanation for why large deficits did not produce high interest rates during the Civil War and World War I. His explanation is also consistent with how consumption expenditures behaved over the sample period 1901–29. Nevertheless, before leaving this section, it is desirable to examine another potential explanation for the unusually low consumption

expenditures during World War I—the price controls then in effect.

According to Frank Taussig (1919), the federal government applied price controls almost exclusively to the goods it was procuring. It did not control the prices of most goods sold at retail. The price controls were therefore essentially hidden excise taxes imposed on the sellers of war matériel to finance hidden expenditures on the war matériel. Since retail prices cleared markets, households may have found goods expensive to buy but not subject to rationing. Therefore, saving probably rose little more than it would have risen without the price controls.

#### IV. The U.S. Experience during World War II

The United States entered World War II in December 1941, and the war lasted until August 1945. Federal spending, which had averaged 12.1 percent of  $GNP$  in 1939–41, rose to an average of 41.3 percent in 1942–45. From an average of 2.6 percent of  $GNP$  in 1939–41, the deficit jumped to an average of 22.7 percent in the next four years.<sup>28</sup> On average, only 11.2 percent of these deficits was monetized, somewhat more than during World War I but much less than during the Civil War.

Using the conventional paradigm, one would predict that interest rates must have risen sharply during World War II. In fact, no such thing happened as can be seen from Figure 3, where I have plotted the commercial paper rate, Moody's Aaa bond rate, and the deficit ratio, labeling them  $CPR$ ,  $AAA$ , and  $DR$ , respectively. It is impossible to see any relationship between  $DR$  and either  $CPR$  or  $AAA$  because the deficit soared during the war while each interest rate followed a path resembling the cardiogram of a rock.

It is not difficult to understand why interest rates were so stable. During the war the Federal Reserve pegged the interest rates on Treasury securities, and interest rates on other securities maintained approximately

<sup>26</sup>One of my referees thinks that  $DR$  may enter my interest rate regressions with negative coefficients because interest rates respond more to the transitory component of  $GR$  than to the permanent component and because these two components are positively correlated. The variable  $DR$  then receives a negative coefficient because it is strongly correlated with the transitory component and because  $GR$  moves more than its transitory component. I have not tested this hypothesis because the univariate methods that I have employed for constructing the permanent and transitory components assume that the two components are orthogonal. To test this hypothesis would require a multivariate model of permanent federal spending, a model that I think would be almost impossible to construct satisfactorily.

<sup>27</sup>Since  $GT$  has an unconditional mean of zero, my proxy for it has a zero sample mean. Homotheticity of household preferences between current and future consumption of nondurables and the service flow from durables would then imply a zero constant term.

<sup>28</sup>Adjusting the World War II figures for inflation as discussed in fn. 4 lowers them on average by about 2.5 percent of  $GNP$ .

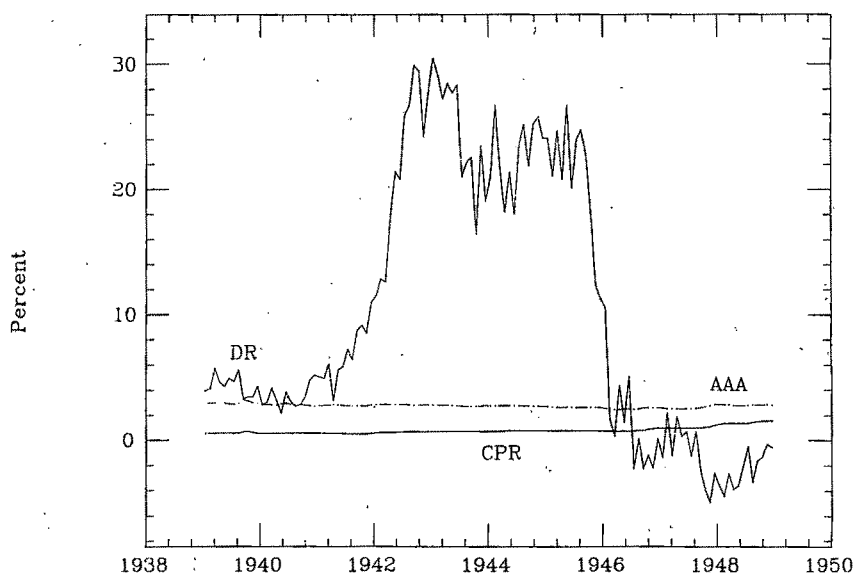


FIGURE 3. U.S. EXPERIENCE DURING WORLD WAR II

constant differentials reflecting differences in default risk, tax treatment, etc. The rates pegged were low enough to create excess demand for goods. Rather than letting the price level rise, the federal government imposed price controls, rationing the available output among eager demanders. Since the government was first in line, it obtained what it sought while consumers and investors obtained whatever the government chose to let them have of what remained. Table 5 shows that the government chose to take about three-fifths from consumption expenditure and about two-fifths from investment expenditure.<sup>29</sup> The government achieved this result by raising the portion of *GNP* absorbed as taxes by 5.9 percentage points, by raising the effective marginal tax rate on corporate profits from an average of 42 percent in 1939–41 to an average of 71.5 per-

cent in 1942–45,<sup>30</sup> and most important by extensive rationing controls.

I cannot directly test whether the large World War II deficits would have produced high interest rates, had interest rates not been pegged and had prices not been controlled. This proposition must necessarily be tested indirectly. I have based my test on the theory outlined below, which shows that there should be a *ceteris paribus* positive association between the deficit and the real money stock if deficits raise desired expenditure and hence unconstrained equilibrium interest rates.

When the government represses inflation and therefore must ration goods, the demand for the real money stock depends not only on real income and the nominal interest rate but also on how restrictive the rationing constraints are. To see why, consider a money demand model of the Baumol-Tobin type. When rationing is in effect, the interest earned by the asset alternative to money does not entitle the recipient to buy anything unless he or she can also obtain ration tickets. If  $R$  is the interest rate on the alternative asset

<sup>29</sup>Robert Gordon (1969) has shown that the figures in Table 5 substantially overstate the effective share of government purchases and understate the effective share of gross total investment. The official measure of government purchases includes, and the official measure of gross total investment excludes, \$45 billion in capital goods that businesses probably knew they would get when the war ended.

<sup>30</sup>See Seater.

TABLE 5—DATA FOR WORLD WAR II<sup>a</sup>

Years	Total Government Purchases/ <i>GNP</i>	Consumption Expenditure/ <i>GNP</i>	Gross Total Investment/ <i>GNP</i>
1939–41	16.3	69.8	13.9
1942–45	42.3	53.8	3.9
Change	26.0	–16.0	–10.0

<sup>a</sup>Shown in percent.

and  $Z$  is the shadow cost of ration tickets, the effective yield on the alternative asset is only  $R/(1+Z)$ ; which is smaller, the more binding rationing constraints are. Economic agents balance off this effective yield on the alternative asset against the transaction costs of shifts between the alternative asset and money. If the transaction costs are largely nonpecuniary—taking the form of inconvenience or, for many currency holders, the risk of going to prison for dealing in the black market—a higher  $Z$  implies little if any increase in transaction costs. Given a lower effective yield on the alternative asset but essentially the same transaction costs, economic agents increase their holdings of money as  $Z$  rises. (See my 1982 paper for more discussion.) I therefore assume that the money demand function takes the form

$$(13) \quad M/P = L\left(\underset{+}{Y}, \underset{-}{R}, \underset{+}{Z}, \underset{+}{UM}\right),$$

where  $M$  is the money stock,  $Y$  is real income,  $R$  is the nominal interest rate,  $Z$  is the shadow price of ration tickets, and  $UM$  is an error term. The function  $L$  is increasing in  $Y$  and  $Z$  and decreasing in  $R$  as indicated by the signs below  $Y$ ,  $R$ , and  $Z$ .

I assume that desired expenditure is given by

$$(14) \quad E = E\left(\underset{-}{Z}, \underset{+}{Y}, \underset{-}{R}, \underset{+}{M/P}, \underset{+}{G}, \underset{+}{D}, \underset{+}{UE}\right),$$

where  $UE$  is an error term. According to equation (14), desired expenditure is decreasing in the shadow price of ration tickets and the nominal interest rate,<sup>31</sup> and is increasing

in income, the real money stock, government spending and the deficit. For  $Z=0$ , equation (14) takes exactly the form that the conventional paradigm assumes. In particular,  $\partial E/\partial D > 0$  because increasing  $D$  raises private disposable income and thus raising desired consumption expenditure and desired total expenditure.

The government issues just enough ration tickets to permit the entire output  $Y$  to be purchased. In equilibrium,  $Z$  assumes the value that equates desired expenditure to output:

$$(15) \quad E(Z, Y, R, M/P, G, D, UE) = Y.$$

Since there were no formal markets in ration tickets during World War II and thus no direct measure of  $Z$ , one cannot estimate equations (13) and (15) individually. I therefore eliminate  $Z$  between them, obtaining the estimable reduced form<sup>32</sup>

$$(16) \quad M/P = F\left(\underset{+}{Y}, \underset{-}{R}, \underset{+}{G}, \underset{+}{D}, \underset{+}{UM}, \underset{+}{UE}\right).$$

I have estimated equation (16), applying ordinary least squares to monthly data that span the period October 1942 to December 1947. In the former month, Congress granted the Roosevelt Administration full power to

<sup>31</sup>Since price controls fix the price level over time,  $R$  is within a constant of the real interest rate.

<sup>32</sup>I assume here that the condition for market stability  $(\partial L/\partial Z)(\partial E/\partial(M/P)) < -\partial E/\partial Z$  holds. I also assume that  $Y$  is predetermined vis-à-vis  $UM$  and  $UE$  for at least a month, the sampling frequency of regressions reported below. Note that pegging  $R$  requires the Federal Reserve always to accommodate  $UM$ , so this assumption is sensible for this error term. It is also sensible for  $UE$  if it takes a month or more for economic agents to alter production and factor supply decisions in response to autonomous changes in expenditures.

set prices; by the latter month, controls had lapsed for the most part, and prices had risen enough to be near equilibrium levels. I assume that  $\ln(M/P)$  is approximately linear in an error term, which is a composite of  $UM$  and  $UZ$ ; and in  $\ln Y$ ,  $R$ ,  $GR$ , and  $DR$ .<sup>33</sup> Pretesting revealed that differencing the reduced form apparently eliminated the serial correlation in the residuals. What I obtained appears below:

$$(17) \quad \Delta \ln(\hat{M}_t/P_t) = .00704 + .325\Delta \ln Y_t \\ (.00315) \quad (.191) \\ - .271\Delta CPR_t + .00355\Delta GR_t - .00246\Delta DR_t, \\ (.134) \quad (.00159) \quad (.00145)$$

$$R^2 = .213, S.E. = .0227, D-W = 1.87.$$

Equation (17) generally confirms the theoretical predictions. The coefficient on  $R$  is negative, the coefficient on  $GR$  is positive, and both are statistically significant at the .05 level. What does not confirm the theory is  $DR$ 's coefficient. Rather than being positive, it is negative though not significantly so at the .05 level. Apparently desired expenditure did not rise with the deficit; if anything, it fell. Therefore, the large deficits during World War II would probably not have produced high interest rates even if interest rates had not been pegged and prices had not been controlled.

The federal government conscripted soldiers during the Civil War, World War I, and World War II. A referee has suggested that conscription may have imposed liquidity constraints on households, reducing consumption expenditure by an empirically important amount. The estimated effects of the deficit ratio on interest rates and desired expenditure may simply reflect a positive correlation of the deficit with conscription.

<sup>33</sup>I used the following proxies for  $M$ ,  $P$ ,  $Y$ ,  $R$ ,  $GR$ , and  $DR$ : the  $M1$  money stock; the Consumer Price Index; personal income deflated by the  $CPI$  (similar results obtain for industrial production); the commercial paper rate; and nominal federal spending and the nominal deficit deflated by the  $CPI$  and trend real  $GNP$ . See the data appendix, which describes and lists the data and provides their sources.

For World War II, I can test this hypothesis since good monthly data are available on military "employment," much of which was drafted. I have fitted a regression of the form (17) adding as an explanatory variable  $MILR$ , the ratio of military employment to the noninstitutional population. The result is

$$(18) \quad \Delta \ln(\hat{M}_t/P_t) = .00738 + .320\Delta \ln Y_t \\ (.00315) \quad (.190) \\ - .269\Delta CPR_t + .00310\Delta GR_t \\ (.133) \quad (.00163) \\ - .00225\Delta DR_t + .00829\Delta MILR_t, \\ (.00145) \quad (.00692)$$

$$R^2 = .233, S.E. = .0227, D-W = 1.89.$$

The coefficient on the deficit ratio remains negative, and the estimated military ratio has the wrong sign to support the hypothesis. This regression is also evidence against the joint hypothesis that patriotism reduced desired consumption expenditures and that it is well proxied by the size of the military.

The conclusions drawn from equation (17) do not change if one uses the ratio of the change in the real market value of gross federal debt to trend real  $GNP$  in place of  $DR$ .<sup>34</sup>

## V. The Postwar Experience

Several researchers have attempted to find an association between nominal interest rates and the U.S. deficit using postwar data. Although a few like Martin Feldstein and Otto Eckstein (1970) have found a weak positive, but statistically significant, association, many others have found none at all.<sup>35</sup> For refer-

<sup>34</sup>Table 7 of W. Michael Cox and Eric Hirschhorn (1983) provides the market value of the gross federal debt over the period between January 1942 and December 1980. Note that the ideal measure of the deficit is just the change in the real market value of federal indebtedness. The "corrected" deficits calculated above are only approximations to this ideal. The data of Cox and Hirschhorn permit one to calculate the ideal without approximation.

<sup>35</sup>See Thomas Sargent (1973).

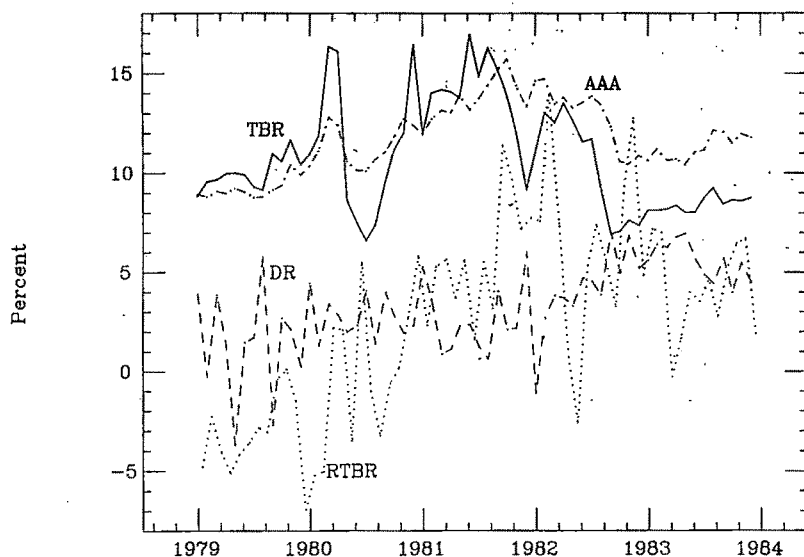


FIGURE 4. U.S. EXPERIENCE SINCE OCTOBER 1979

ences to this literature and the best article in my opinion on the subject, see Charles Plosser (1982).

Most of these studies have three important weaknesses. First, the Federal Reserve stabilized interest rates over most of the postwar period, perhaps hiding the true relationship. Second, prior to the 1980's, the deficit was rarely large and did not vary much. When it did vary, it generally did so because of the business cycle. Third, the endogeneity of federal spending, deficits, and the money stock has often been ignored or poorly handled.

In this section, I investigate the relationship between interest rates and deficits in the period between October 1979 and December 1983. During this period, the Federal Reserve largely freed interest rates to seek their own levels. Furthermore, deficits became unusually large (in part because of the Reagan Administration tax cut);<sup>36</sup> they varied considerably; and much of that variance reflected such exogenous events as the defense

buildup, three cuts in personal income tax rates, the acceleration of depreciation allowances, and a tax hike. Therefore, if large deficits do produce high interest rates and one is ever to show that they do, one should be able to do so with these data.

Figure 4 plots the deficit ratio, the three-month Treasury bill rate, Moody's Aaa bond rate, and the *ex post* real one-month Treasury bill rate over the sample period, labeling them *DR*, *TBR*, *AAA*, and *RTBR*.<sup>37</sup> There is no obvious positive association between *DR* and either *TBR*, *AAA*, or *RTBR*. If anything, the plot suggests a negative association, which may, however, be misleading since the deficit was a highly endogenous variable during this period, rising in recessions and falling in booms. For this reason, it is important to estimate the relationship between interest rates and deficits, taking proper account of the endogeneity of the deficit.

I have fitted the three-month Treasury bill rate, Moody's Aaa bond rate, and the real one-month Treasury bill rate to constant terms, time trends, and current and past

<sup>36</sup>Barro (1983), however, has adduced evidence that deficits have not been unusually large, given the large amount of slack in the economy and the size of the government.

<sup>37</sup>*RTBR* is the one-month Treasury bill rate minus the *CPI* inflation rate over the next month.

TABLE 6—REGRESSIONS FOR THE PERIOD SINCE OCTOBER 1979<sup>a</sup>

	(6.1)	(6.2)	(6.3)	(6.4)	(6.5)	(6.6)
	Dependent Variable					
	<i>TBR</i>	<i>AAA</i>	<i>RTBR</i>	<i>TBR</i>	<i>AAA</i>	<i>RTBR</i>
Constant	2.7 (49.2)	23.6 (23.6)	101.3 (64.0)	-38.0 (39.5)	14.0 (14.2)	73.8 (38.8)
Trend	.096 (.111)	.0843 (.0530)	.039 (.144)	.180 (.113)	.0885 (.0406)	.079 (.111)
$GR_t$	1.52 (1.78)	.841 (.855)	-0.48 (2.32)	2.37 (1.21)	.912 (.435)	-0.40 (1.19)
$GR_{t-1}$	-0.82 (0.85)	-.421 (.407)	-0.24 (1.10)			
$GR_{t-2}$	0.37 (0.56)	.099 (.202)	.128 (.269)			
$DR_t$	-2.76 (2.15)	-1.47 (1.03)	0.19 (2.80)	-3.63 (1.25)	-1.40 (0.45)	-0.30 (1.23)
$DR_{t-1}$	0.56 (0.75)	0.21 (0.36)	-0.36 (0.98)			
$DR_{t-2}$	-0.18 (0.54)	0.10 (0.26)	0.38 (0.70)			
$MR_t$	-9.1 (9.0)	-1.58 (4.33)	-14.8 (11.7)	0.41 (1.40)	-1.36 (0.50)	-4.54 (1.38)
$MR_{t-1}$	11.8 (10.6)	1.78 (5.06)	16.1 (13.7)			
$MR_{t-2}$	-3.4 (4.0)	-1.69 (1.91)	-6.1 (5.2)			
<i>S.E.</i>	2.81	1.35	3.65	3.57	1.28	3.51
<i>D-W</i>	1.88	1.83	1.44	1.74	1.72	1.70

<sup>a</sup>Standard errors are shown in parentheses.

federal spending ratios, deficit ratios, and money ratios, using monthly data that span the period October 1979 to December 1983. I have treated the current values of the ratios as endogenous, using two-stage least squares.<sup>38</sup> Table 6 reports some regressions

that I have fitted. Regressions (6.1)–(6.3) include two lags on each right-hand side variable. In none of these regressions do the current and lagged values of the deficit have positive coefficients statistically significant at any reasonable level. Rather, they generally have negative, albeit insignificant, ones. Dropping the least significant variables sequentially, I eventually obtained regressions (6.4)–(6.6), the best fitting regressions that

<sup>38</sup>As instruments, I used the ratio of real federal spending by the Defense Department to trend real *GNP*; the ratio of federal employment to the population; two dummy variables; the discount rate of the New York Federal Reserve Bank; and bank borrowing from the Federal Reserve deflated by the *CPI* lagged one month and trend real *GNP*. One dummy variable was 0 from October 1979 to September 1981, 5 from October 1981 to June 1982, 15 from July 1982 until June 1983, and 25 from July 1983 until December 1983. These figures are roughly the percentage amounts that the Economic Recovery Act cut personal income tax rates. The other dummy variable was 0 from October 1979 to December 1982 and 1 from January 1983 to December 1983, when the Tax Equity and Fiscal Responsibility Act and the Highway Revenue Act were in effect. The federal defense spending ratio, the federal employment ratio, the two dummy variables, and the discount rate qualify as

an exogenous variables. Furthermore, over the sample period, the operating procedures of the Federal Reserve make bank borrowing a good candidate for exogeneity. Lagged reserve requirements essentially predetermined total reserves. The Federal Reserve had to supply the predetermined quantity of total reserves but could, through its control over nonborrowed reserves, fix the amount that the banks had to borrow in order to satisfy their reserve requirements, thereby influencing their portfolio decisions and the demand deposits that they supplied. Because the current price level might be affected by current Federal Reserve policy, I have deflated by the previous month's *CPI*.



retain *GR*, *DR*, and *MR* in some form. Each regression keeps only the contemporaneous terms *GR* and *DR*. In regressions (6.4) and (6.5), *GR* has positive coefficients that are statistically significant at the .05 level. These positive coefficients are consistent with the conventional paradigm. In contrast, *DR*'s uniformly negative coefficients are inconsistent with the conventional paradigm, especially the significant coefficients in regressions (6.4) and (6.5). I conclude from these results, not that the large deficits in 1982 and 1983 lowered interest rates, but rather that there is no evidence that they produced the high interest rates that have prevailed since October 1979.

Why have interest rates responded so little to deficits in the postwar period? Several studies, the best of which in my opinion is Roger Kormendi (1983), suggest that changes in the deficit in the postwar period have been offset by essentially equal changes in private saving, thereby removing the need for interest rates to change. See Kormendi for references to this literature.

## VI. An Explanation

What needs to be explained is why, in over a century of U.S. history, large deficits have never been associated with high interest rates, why the large increase in the deficit during World War I was associated with a large almost equal increase in private saving, why private saving has moved in the postwar period to offset changes in the deficit, and why large deficits apparently did not produce high aggregate demand. The explanation that seems most consistent with these observations is Barro's.<sup>39</sup>

Barro has argued that it may be optimal for households to react to an increased deficit by increasing their saving by an equal amount. Consequently, neither aggregate demand nor interest rates may rise. Households

will so react if capital markets are perfect, if they understand the intertemporal budget constraints they face, and if they have operative altruistic intergenerational transfer motives. They will then know that the current deficit equals in present value the taxes to service the extra debt. An increase in the deficit will therefore entail an equal increase in saving, which will just suffice to pay the extra future taxes levied on present households and subsequent generations for whom they care.

Macroeconomists have hesitated to accept Barro's argument, in part because they have doubted that households can foresee the higher future taxes implied by a larger deficit, and in part because they have doubted that households have altruistic intergenerational transfer motives. The assumption of accurate foresight of future tax liabilities does indeed seem implausible. One should, however, judge the utility of an assumption primarily by its predictive and explanatory power and not by its realism. The phenomena detailed above are consistent with accurate foresight but not with the very limited foresight assumed in the conventional paradigm.

Unlike the assumption of accurate foresight, the assumption of altruistic intergenerational transfer motives may actually turn out to be realistic. Laurence Kotlikoff and Lawrence Summers (1981) have presented evidence that the bulk of U.S. saving takes the form of intergenerational transfers.<sup>40</sup> Furthermore, casual observation suggests that households spend a great deal on their children—incomprehensible behavior if households actually have the preferences assumed in many macro models. Ultimately, however, the proof of this assumption lies in its predictive and explanatory power.

I have found for all of my samples that increasing the deficit ratio while keeping the federal spending ratio constant (i.e., cutting current average and marginal tax rates) actu-

<sup>39</sup>This explanation antedates Barro. In fact, it is generally termed *Ricardian equivalence* because Ricardo advanced it in his "Funding System." See Piero Sraffa (1951). Martin Bailey (1971) has also advanced this explanation.

<sup>40</sup>Franco Modigliani (1983) has presented some evidence that the figures of Kotlikoff and Summers are inflated. In any case, intergenerational transfers are probably empirically important.

ally lowers interest rates, generally by a statistically significant amount. I have never emphasized this finding, being content to show how little evidence there is that increased deficits raise interest rates. If, however, one should decide to take this finding seriously, what explanation can one offer? One possibility is that the lower tax rates associated with a larger deficit raise the after-tax real return to saving, raising saving appreciably as Boskin has argued. Therefore, private saving rises not only by enough to pay all future tax liabilities à la Barro, but also by an additional amount, thereby lowering aggregate demand and hence interest rates.

### VII. Conclusions

Economists like to think of economics as a science. In a science, however, repeated contradictions of a paradigm lead to its abandonment if there is any sensible alternative. One paradigm in economics implies that large deficits produce high interest rates. This paradigm is not supported by the facts. In over a century of U.S. history, large deficits have never been associated with high interest rates. Even the postwar periods separately offer no support for a positive association between deficits and interest rates. Indeed, the evidence more strongly supports a negative association than a positive one. Therefore, it seems that economists should be looking for an alternative paradigm to replace their current one. The previous section suggested that this paradigm should contain elements of Barro's model and should allow saving to increase with its real after-tax return.

This paper should not be taken to support deficit spending. In my view, which is the same as Barro's (1979), the government ought to run whatever deficit or surplus is necessary to flatten the expected future time profiles of its marginal tax rates. Deficits reduce welfare if they lead to an expected upward trend in marginal tax rates. Conversely, surpluses reduce welfare if they lead to an expected downward trend. Therefore, concern about current deficits should focus on whether federal spending will eventually be

cut enough that current marginal tax rates will finance it, or whether marginal tax rates will have to rise. Concern should not focus on what deficits do to interest rates, capital accumulation, or economic growth, for there is precious little evidence that deficits affect these variables.

### REFERENCES

- Bailey, Martin J., *National Income and the Price Level*, 2d ed., New York: McGraw-Hill, 1971.
- Barro, Robert J., "Are Government Bonds Net Wealth?," *Journal of Political Economy*, November/December 1974, 82, 1095-117.
- , "On the Determination of the Public Debt," *Journal of Political Economy*, October 1979, 87, 940-71.
- , "Output Effects of Government Purchases," *Journal of Political Economy*, December 1981, 89, 1086-121.
- , "The Behavior of U.S. Deficits," unpublished, University of Chicago, March 1983.
- Boskin, Michael J., "Taxation, Saving, and the Rate of Interest," *Journal of Political Economy*, April 1978, Part II, 86, S3-S27.
- Cox, W. Michael and Hirschhorn, Eric, "The Market Value of the U.S. Government Debt; Monthly, 1942-1980," *Journal of Monetary Economics*, March 1983, 11, 261-72.
- Engerman, Stanley L., "The Economic Impact of the Civil War," *Explorations in Entrepreneurial History*, Spring/Summer 1966, 3, 176-99.
- Evans, Paul, "The Effects of General Price Controls in the U.S. During World War II," *Journal of Political Economy*, October 1982, 90, 944-66.
- Feldstein, Martin, "Signs of Recovery," *Economist*, June 11, 1983, 287, 43-48.
- and Eckstein, Otto, "The Fundamental Determinants of the Interest Rate," *Review of Economics and Statistics*, November 1970, 52, 363-75.
- Friedman, Milton and Schwartz, Anna J., *A Monetary History of the United States, 1867-1960*, Princeton: Princeton University Press, 1963.

- Gordon, Robert J., "\$45 Billion of U.S. Private Investment Has Been Mislaid," *American Economic Review*, June 1969, 59, 221-38.
- Homer, Sidney, *A History of Interest Rates*, New Brunswick: Rutgers University Press, 1963.
- Kormendi, Roger C., "Government Debt, Government Spending, and Private Sector Behavior," *American Economic Review*, December 1983, 73, 994-1010.
- Kotlikoff, Laurence J., and Summers, Lawrence H., "The Role of Intergenerational Transfers in Aggregate Capital Accumulation," *Journal of Political Economy*, August 1981, 89, 706-32.
- Mitchell, Wesley C., *A History of the Greenbacks*, Chicago: University of Chicago Press, 1903.
- Modigliani, Franco, "The Life Cycle Hypothesis and National Wealth—A Rehabilitation," unpublished, Massachusetts Institute of Technology, January 1983.
- Plosser, Charles I., "Government Financing Decisions and Asset Returns," *Journal of Monetary Economics*, May 1982, 9, 325-52.
- Roll, Richard, "Interest Rates and Price Expectations during the Civil War," *Journal of Economic History*, June 1972, 32, 476-98.
- Sargent, Thomas J., "The Fundamental Determinants of the Interest Rate: A Comment," *Review of Economics and Statistics*, August 1973, 15, 391-93.
- Seater, John J., "Marginal Federal Personal and Corporate Income Tax Rates in the U.S., 1909-1975," *Journal of Monetary Economics*, November 1982, 10, 361-82.
- Sraffa, Piero, *The Works and Correspondence of David Ricardo*, Vol. 4, Cambridge: Cambridge University Press, 1951.
- Taussig, Frank W., "Price-Fixing as Seen by a Price-Fixer," *Quarterly Journal of Economics*, February 1919, 33, 205-41.

# Transaction Costs in the Theory of Unemployment

By PETER HOWITT\*

This paper addresses the problem of accounting theoretically for the persistence of large-scale unemployment. It briefly reviews some of the well-known shortcomings of the Keynesian account based upon sticky wages and prices and upon Robert Clower's (1965) concept of effective demand, arguing that these shortcomings can be seen as a failure explicitly to integrate transaction costs into the theory of unemployment. The main constructive contribution of the paper is an example of a simple macro model, along the lines of Robert Barro and Herschel Grossman (1971), in which transaction costs are made explicit. An essential part of this example is an externality of the sort suggested by Peter Diamond (1982). The example provides an account of persistent unemployment with many Keynesian features, yet without some of the Keynesian shortcomings. In particular it assumes perfect wage and price flexibility.

## I. The Keynesian Account

For future reference, I sketch the standard Keynesian account in terms of Figure 1, reproduced in essence from Barro and Grossman (p. 86). Both the output and labor market are in excess supply at a real wage equal to its general equilibrium value,  $w^*$ . The supply of labor equals its notional demand  $n^*$ . But the effective demand, as given by the schedule  $ABn'$ , is only  $n'$ , because collectively the firms find themselves unable to sell any more output than can be produced by this amount. Thus the amount  $n^* - n'$  of unemployment exists.

\*University of Western Ontario, London, Ontario N6A 5C2, Canada. This paper has benefited from the helpful comments and criticisms of Joel Fried, Tom Kompas, Richard Manning, John McMillan, Tom Rymes, Dan Usher, and John Vanderkamp; none of whom is responsible for any shortcomings.

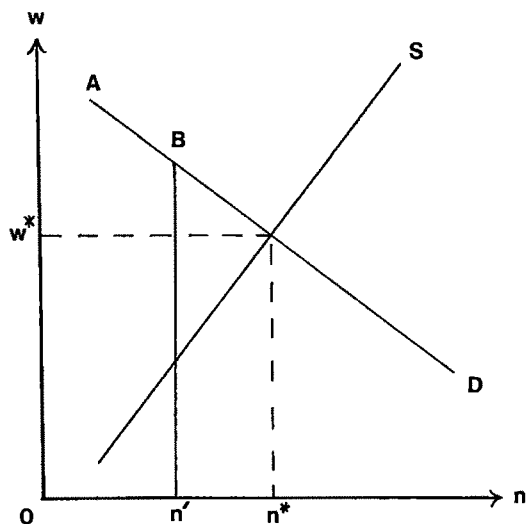


FIGURE 1

Over time the money wage and the price level will fall in response to excess supplies. This will generally raise the aggregate demand for goods, shifting the vertical portion of the effective demand for labor schedule to the right, and thus increasing employment. Less than full employment will persist if the process of deflation is slow, or if the effect of deflation on aggregate demand is small (or in the wrong direction). Involuntary unemployment will also persist, with  $w$  exceeding the supply price of labor, unless wages fall more rapidly than prices, driving real wages down to the point where effective demand equals notional supply.

This account has many virtues. It is consistent with the widely observed lags in wages and prices behind the business cycle with the countercyclical pattern of layoffs, and with the absence of any pronounced cyclical pattern to real wages. The assumption of quantity rationing underlying Clower's distinction between notional and effective demands also makes the approach consistent with the apparently involuntary nature of large-scale

unemployment, and with the observation that the typical business firm would usually prefer to increase its sales at the current price.

There are, however, some well-known shortcomings. As Barro (1977) has pointed out, the account gives no clear explanation of why the mutually advantageous gains from trade implied by the fact that the supply price of labor at  $n'$  is less than its marginal product remain persistently unexploited. Defenders of the approach often take this failure as a logical consequence of wage-price stickiness. But Barro points out that such stickiness can be explained by theories which predict that all potential gains from trade will be exhausted. For example, implicit contracts might specify efficient quantities of goods and services to be traded, along with money payments that are set in fixed ratio to the corresponding quantities in order to satisfy an independent condition of optimal risk sharing.<sup>1</sup> Thus something in addition to price stickiness must be adduced in order to explain the market failure in Figure 1.

That additional something is usually presented by defenders of theories of non-clearing markets as being the limitations imposed by transaction costs upon people's ability to communicate offers to buy and sell.<sup>2</sup> The stickiness of prices in Keynesian economics results not from a full set of pre-arranged contractual agreements but from an institutional assumption that prices must be set and advertised by agents (often personified collectively by the "auctioneer") before there is time for everyone to agree upon what quantities to trade. These costs are what prevent the exhaustion of all gains from trade. But the main problem with this line of defense is that the transaction costs are

not integrated explicitly into the theory of unemployment. They do not appear in the decision problems faced by agents, except sometimes as part of an explanation of the demand for money.

The problem is manifested in a difficulty with the concept of effective demand. Specifically, why doesn't the firm in Figure 1 try to remove its sales constraint by undercutting its rivals' price? If it made the conjecture most compatible with the price-taking assumptions of the Barro-Grossman model, it would see itself as being able to supply anywhere up to the entire market demand at anything less than the going price. In this case there would be no reason why the demand for labor should not be determined by the notional schedule in Figure 1, because a rational firm would ignore any constraint that could be removed by a negligible price reduction.<sup>3</sup> The concept of effective demand for labor would be redundant and there would be no unemployment in Figure 1. This difficulty might be resolved by invoking any of a number of costs of price adjustment. But the costs are absent from the standard Keynesian account.<sup>4</sup>

The problem is also revealed by recalling that, even if we accept the effective demand schedule  $ABn'$  as defining the relevant demand for labor, unemployment can persist in Figure 1 only if the money wage rate falls slowly relative to the price level. According to the usual textbook interpretation of the tatonnement process, this means that people remain unemployed only because they are slow to offer their services at less than the going wage. Similarly, the "contract" inter-

<sup>3</sup>This problem was pointed out by Don Patinkin (1965, p. 323, fn. 9).

<sup>4</sup>See, however, the attempt by Katsuhito Iwai (1981) to develop a general theory of nonclearing markets upon a microfoundation that makes explicit the institutional restrictions according to which firms set their prices. See also Geoffrey Woglom's (1982) macro model, in which the failure of firms to engage in competitive price reductions whenever they face prices in excess of marginal cost is based upon Joseph Stiglitz's (1979) analysis of markets with costly communication. Woglom's analysis is similar to the conjectural equilibrium analysis of Frank Hahn (1978), which provides another possible approach to resolving this difficulty.

<sup>1</sup>The model of Costas Azariadis (1975) actually implies a fixed real wage rather than the (temporarily) fixed nominal wage of Keynesian theory. However, Bennett McCallum (1978) has pointed out that the important qualitative implications of a natural-rate model with rational expectations and the Lucas aggregate-supply schedule, from which presumably all Keynesian elements of market failure are absent, would go through unaffected even if nominal and prices were preset by contractual agreements.

<sup>2</sup>See, for example, Axel Leijonhufvud (1968), Arthur Okun (1981), and David Laidler (1982).

pretation of Stanley Fischer (1977) and others requires the assumption that unemployed workers do not attempt to underbid anyone working on an unexpired contract (see Fischer, p. 198, fn. 17). This is not a convincing description of "involuntary" unemployment. As Axel Leijonhufvud stressed in his well-known critique, Keynesian economics thus agrees with the pre-Keynesian diagnosis that people are unemployed only because they are asking too much. One might defend Keynesian economics from this critique by arguing that the slowness with which the unemployed offer their services at less than the going wage is a theoretical proxy for the time cost of communicating and forming new matches in the labor market. But again these costs are absent from the account.

## II. Transaction Costs and Thin Markets

The preceding discussion attributes many of the shortcomings of the Keynesian account to the absence of explicit transaction costs, broadly interpreted to include costs of communication. It is a commonplace that such costs are higher the thinner the market; that the per unit cost of transacting depends inversely upon the amount of activity in the market. This observation refers not to the well-known economies of scale from lumpy set-up costs (for example, William Baumol, 1952), but to an externality whereby one agent's trading costs are reduced by having agents on the other side of the market devote more resources to trading.

The observation usually refers to cross-market comparisons, as between over-the-counter and regularly listed stock transactions, but it could equally well apply to intertemporal comparisons within any given market. For example, a decrease in the demand for labor makes it harder to find a job as well as reducing the wage, because potential employers advertise less and become less willing to arrange interviews, read job applications, return calls, and so forth.<sup>5</sup> From the

potential employer's point of view, the labor market becomes less thin when the number of unemployed workers searching for jobs increases and the cost of finding suitable potential recruits thereby decreases. In markets for consumer durables, when demand falls many people stop reading advertisements, stop visiting sellers, and generally become more difficult for sellers to contact. The phenomenon is especially marked in the housing market, where a decrease in demand not only reduces the market price but also increases the expected waiting time required to sell at that price.

This externality works mainly through the cost of communication. The extra cost of dealing in a thin market is primarily that of identifying, contacting, and negotiating with a suitable trading partner. This suggests that the externality is particularly difficult to internalize. Such internalization would generally require some kind of collective agreement to coordinate the activities of potential traders on both sides of the market. But the agreements themselves would require communications whose costs are subject to the externality. Once two potential trading partners have contacted one another, it is generally too late for them to agree to an arrangement to share the costs of contacting in such a way as to induce the efficient amount of contacting activity by each side.<sup>6</sup>

Some internalization obviously is accomplished by intermediaries in the labor market, such as employment agencies, university placement services, newspapers, and trade associations; as well as by retail and wholesale firms, auction houses, jobbers, specialist traders, financial intermediaries, real estate agents, etc. in other markets. But the scope of such intermediation in labor markets is limited by the problems of heterogeneity and

<sup>5</sup>How this can happen and why searching workers are affected by this change in recruiting behavior independently of its effects on wages is described graphically by Okun's analysis of "No-Help-Wanted signs" (pp. 56-61).

<sup>6</sup>Such arrangements are analyzed by Dale Mortenson (1982). As he points out, they are most likely to occur in markets intermediated by brokers who have dealt with each other in the past, either directly or through their common membership in some organized institution, and have thus had the occasion to reach some understanding prior to their current contacting activities. The brokers also need to have an expectation that their mutual dealings will continue in the future in order to have an incentive to adhere *ex post* to agreements that do not have the force of law.

asymmetry of information, which tend to render large-scale intermediation uneconomical. Even in other markets there is no reason to think *a priori* that intermediation eliminates all substantial externalities.

An example of a formal model of this kind of externality was provided by Robert Jones (1976), who based his explanation of the emergence of monetary exchange upon a search model in which the expected time required to contact a trading partner depended inversely upon the number of such potential partners actively in the market.

Diamond has shown, using this kind of search model, how the externality can explain "low-level" equilibria in very special models, which are suggestive of fixed price excess-supply equilibria but without the fixed prices. In these models, a widespread expectation of high costs of contacting trading partners can be self-fulfilling. It will discourage production, thereby resulting in a low volume of trade, and thereby bringing about the expected high cost by thinning out markets.

### III. An Example

The example of this section can be seen as recasting into more familiar macroeconomic terms the basic results of Diamond.<sup>7</sup> Rather than focus exclusively upon one of the many informational, logistical, institutional or strategic factors underlying transaction costs I shall model such costs in the less explicit but more general manner of such writers as Frank Hahn (1971), and Jürg Niehans (1971).

<sup>7</sup>The differences between this model and Diamond's are that: (a) Diamond's model has essentially one kind of agent and one kind of good, rather than firms and households and labor and output; (b) his production technology is described by an exogenous Poisson process rather than a usual short-run production function; (c) his model has no congestion externalities; (d) he describes unemployment as waiting for a production opportunity and employment as searching for a trading partner, which on the face of it seems to get it the wrong way around, whereas this model has more conventional descriptions; (e) his model is explicitly dynamic whereas this one is static; and (f) his model is more explicit about the institutional arrangements underlying transaction costs.

Specifically, assume that traders are convened by an auctioneer (or a set of specialist auctioneers) able to find market-clearing wages and prices at no cost, but unable to arrange the trades costlessly. There are unspecified trading institutions in place that reduce but do not eliminate the trading costs faced by households and firms. In addition to the usual budget constraint, there is a transaction-cost constraint requiring each trader to use up resources in order to execute his planned transactions.

There are only two markets, labor and output; and two types of traders, identical households and identical firms. Let  $\bar{y}$ ,  $\bar{s}_y$ , and  $\bar{b}_y$  denote the quantity of output traded in the market, the amount of output used up by all firms in selling output (their "marketing effort"), and the amount of output used up by households in buying output (their "buying effort"), all measured as aggregate quantities per firm. Each firm takes these quantities as given and faces the transaction-cost constraints:

$$(1) \quad s_y = \bar{\sigma}(\bar{s}_y, \bar{b}_y) y,$$

where  $s_y$  is his own marketing effort,  $y$  is the quantity he plans to sell, and <sup>8</sup>

$$(2) \quad \bar{\sigma} > 0, \quad \bar{\sigma}_1 > 0, \quad \bar{\sigma}_2 < 0.$$

Equation (1) asserts that his selling cost must be incurred in the form of output used up, and that this cost is proportional to the volume of his transactions.<sup>9</sup> The negative dependency upon households' buying effort asserted in (2) represents the externality described in the previous section. The positive dependency upon the marketing effort of the firm's rivals represents an external diseconomy that might reasonably be supposed to

<sup>8</sup>All functions are assumed to be smooth. Partial derivatives are denoted by subscripts. Unless otherwise indicated the domain of each function introduced consists of all strictly positive values of its arguments, and unqualified statements like (2) are understood to hold over the entire domain of the functions involved. All prices and quantities are understood to be strictly positive unless otherwise indicated.

<sup>9</sup>This assumes away the important phenomenon of set-up costs.

counteract that external economy. For example, more advertising and product promotion by rivals might require the firm to increase its own efforts to avoid losing customers.<sup>10</sup>

Averaging (1) across all firms yields

$$(3) \quad \bar{s}_y = \bar{\sigma}(\bar{s}_y, \bar{b}_y) \bar{y}.$$

Assume that households' transaction-cost constraints require a buying effort proportional to the quantity bought. Thus  $\bar{b}_y = \beta_y \bar{y}$  for some constant  $\beta_y \in (0, 1)$ . By specifying  $\beta_y$  as a constant in this example, I am ignoring the external economy conferred upon buyers when firms spent more upon marketing. Substituting for  $\bar{b}_y$  in (3) produces the equation:

$$(4) \quad \bar{s}_y = \bar{\sigma}(\bar{s}_y, \beta_y \bar{y}) \cdot \bar{y}.$$

Assume that for any  $\bar{y} > 0$  there is some  $\bar{s}_y$  satisfying (4), and that

$$(5) \quad \bar{\sigma}_1(\bar{s}_y, \beta_y \bar{y}) \cdot \bar{y} < 1 \quad \text{for all } (\bar{s}_y, \bar{y}).$$

Inequality (5) asserts that a firm does not have to match an increase in its rivals' marketing effort in order to maintain constant sales. These two assumptions imply that (4) can be expressed as:  $\bar{s}_y = \bar{s}_y(\bar{y})$ , with

$$(6) \quad \bar{s}_y'(\bar{y}) = (\bar{\sigma} + \beta_y \bar{y} \bar{\sigma}_2) / (1 - \bar{\sigma}_1 \cdot \bar{y}).$$

Thus the per unit selling cost faced by each firm will be

$$(7) \quad \sigma(\bar{y}) \equiv \frac{\bar{s}_y(\bar{y})}{\bar{y}} \equiv \bar{\sigma}(\bar{s}_y(\bar{y}), \beta_y \bar{y}).$$

Assume that the external economy dominates the congestion externality, in the sense that a proportional increase in both marketing efforts and buying efforts will reduce the per unit selling cost:

$$(8) \quad d\bar{\sigma}(\theta s, \theta b)/d\theta < 0.$$

<sup>10</sup> Such diseconomies are made explicit in the context of hiring labor, rather than selling output, by the analysis of myself and Preston McAfee (1983).

Then an increase in the market quantity will cause a decrease in the per unit cost of selling:

$$(9) \quad \begin{aligned} \sigma'(\bar{y}) &= (1/\bar{y}) (\bar{s}_y'(\bar{y}) - \bar{s}_y(\bar{y})/\bar{y}) \\ &= (1/\bar{y}) (\bar{y} \bar{\sigma} \bar{\sigma}_1 + \bar{y} \beta_y \bar{\sigma}_2) / (1 - \bar{y} \bar{\sigma}_1) \\ &= (1/\bar{y}) (1/(1 - \bar{y} \bar{\sigma}_1)) \\ &\quad \cdot \frac{d}{d\theta} \bar{\sigma}(\theta \bar{s}_y(\bar{y}), \theta \beta_y \bar{y})|_{\theta=1} < 0, \end{aligned}$$

(from equations (5)–(8)).

Finally, assume that

$$(10) \quad 1 + \bar{\sigma} + \bar{y}(\beta_y \bar{\sigma}_2 - \bar{\sigma}_1)|_{(\bar{s}_y, \bar{b}_y) = (\bar{s}_y(\bar{y}), \beta_y \bar{y})} > 0.$$

Assumption (10) is hard to interpret directly. Obviously it implies a limitation on the extent of externalities, since  $\beta_y \bar{\sigma}_2 - \bar{\sigma}_1 < 0$ . It would be implied, given assumption (5), if we assumed that an increase in market demand, accompanied by a corresponding increase in buying effort by households, would not generate so large an external benefit on firms that they could meet this increase without an increase in selling effort; that is, if we assumed that  $\bar{s}_y' > 0$ . For, in that case, (5) and (6) would imply that  $1 + \bar{\sigma} + \bar{y}(\beta_y \bar{\sigma}_2 - \bar{\sigma}_1) = (1 - \bar{y} \bar{\sigma}_1) + (\bar{\sigma} + \bar{y} \beta_y \bar{\sigma}_2) = (1 - \bar{y} \bar{\sigma}_1)(1 + \bar{s}_y'(\bar{y})) > 0$ .

It follows from (5), (9) and (10) that

$$(11) \quad 1 + \sigma(\bar{y}) + \bar{y} \sigma'(\bar{y}) = \frac{1 + \sigma + \bar{y}(\beta_y \bar{\sigma}_2 - \bar{\sigma}_1)}{1 - \bar{y} \bar{\sigma}_1} > 0.$$

An example satisfying all the above conditions is the function:

$$(12) \quad \bar{\sigma}(s, b) \equiv (1 + \beta_y s/b) e^{-(\mu/\beta_y)b}, \quad \mu > 0$$

which yields

$$(13) \quad \sigma(\bar{y}) \equiv e^{-\mu \bar{y}} / (1 - e^{-\mu \bar{y}}).$$



An analogous treatment of the costs of buying and selling labor leads to the per unit cost of selling,  $\tau(\bar{n})$ , where  $\bar{n}$  is the quantity of labor services traded in the market, and  $\tau$  is measured in units of labor. Assume that all transaction costs in the labor market are incurred in the form of labor services used up. I shall interpret unemployment as labor services used up in the selling of labor services. The market quantity of unemployment is  $\bar{s}_n = \tau(\bar{n})\bar{n}$ , and the rate of unemployment is the fraction of all labor services used in selling labor:<sup>11</sup>

$$(14) \quad u(\bar{n}) \equiv \frac{\bar{s}_n}{\bar{s}_n + \bar{n}} = \frac{\tau(\bar{n})}{1 + \tau(\bar{n})}.$$

As with  $\sigma$ , the per-unit cost of selling labor satisfies

$$(15) \quad \tau(\bar{n}) > 0, \quad \tau'(\bar{n}) < 0, \\ 1 + \tau(\bar{n}) + \bar{n}\tau'(\bar{n}) > 0.$$

Therefore the rate of unemployment and the level of employment are inversely related:

$$(16) \quad u'(\bar{n}) = \tau'(\bar{n})/(1 + \tau(\bar{n}))^2 < 0.$$

The cost of buying  $n$  units of labor is  $b_n = \beta_n n$ , for some constant  $\beta_n \in (0, 1)$ .

To sell  $y$  a firm must produce  $y(1 + \sigma(\bar{y}))$ . If it hires  $n$  it can use  $n(1 - \beta_n)$  in production. It therefore chooses  $n$  to maximize its

profit:

$$f(n(1 - \beta_n))/(1 + \sigma(\bar{y})) - wn,$$

where  $w$  is the real wage and  $f$  a production satisfying

$$(17) \quad f'(x) > 0; f''(x) < 0 \quad \text{for all } x > 0;$$

$$\lim_{x \rightarrow 0} (f(x), f'(x)) = (0, \infty);$$

$$\lim_{x \rightarrow \infty} (f(x), f'(x)) = (\infty, 0).$$

Given any  $(w, \bar{y})$  the firm's demand for labor is uniquely determined by the first-order condition:

$$(18) \quad w = \frac{1 - \beta_n}{1 + \sigma(\bar{y})} f'(n(1 - \beta_n)),$$

which can be solved for the demand-for-labor function  $n^d(w, \bar{y})$ , with

$$(19) \quad n^d > 0, \quad n_1^d < 0, \quad n_2^d > 0.$$

Note that this demand function depends not only upon the real wage, but also upon the realized quantity of aggregate demand,  $\bar{y}$ . This is because, in order to formulate its plans, the firm must know the per unit cost of selling output, which depends upon aggregate demand; as in the Barro-Grossman model, the firm must receive quantity signals as well as price signals.

As Clower pointed out, this dependency of demand upon realized quantities is what distinguished Keynes' concept of effective demand from the Walrasian concept of notional demand. But the present derivation of the effective demand for labor does not require firms to make unrealistically pessimistic conjectures. Although they take the quantity demand per firm as given, they see themselves as able to sell more if they wish at the going price, as long as they are willing to pay the marketing cost. An increase in aggregate demand for output raises the demand for labor even with no change in the real wage, not because it raises the maximum

<sup>11</sup>Interpreting (14) as the rate of unemployment requires one to identify  $\bar{n}$  as not only the quantity traded but also the quantity employed. This is required again when  $\bar{n}$  is used as the argument of a production function. I am thus abstracting from one of the most important implications of Okun's "Toll" model, namely that the costly trades in the labor market are new matches. Thus transaction costs would perhaps better be described as a function of the rate of increase in employment. This timeless model might be interpreted as describing stationary quantities chosen by households and firms who realize that more employment will, in the stationary state, imply more turnover (because, say, job separations occur at some exogenous proportional rate), and hence more new matches to be formed. Obviously a fuller treatment of time is required for an adequate treatment of this issue.

amount a firm can sell, but because it makes any given amount of output easier to sell.<sup>12</sup>

Each household's utility function has the form  $U(z, m) - c(l)$ , where  $z$  is consumption,  $m$  is demand for real money balances, and  $l$  is supply of labor. Assume that  $U$  is homogeneous of degree one with indifference curves that do not touch the axes. Assume also that there is an upper limit  $\bar{l} > 0$  to each household's potential labor supply and that the cost function  $c$  is defined over the interval  $[0, \bar{l}]$ , with

$$(20) \quad c'(l) > 0 \text{ and } c''(l) > 0 \\ \text{for all } l \in (0, \bar{l}); \\ c'(0) = 0 \text{ and } \lim_{l \rightarrow \bar{l}} c'(l) = \infty.$$

The assumptions of homogeneity and additivity eliminate income effects from labor supply and permit a diagrammatic analysis similar to Figure 1.

The household that buys  $y$  and sells  $n$  gets to consume  $y(1 - \beta_y)$  and must supply  $n(1 + \tau(\bar{n}))$ . Thus the household's decision problem is to choose  $y$ ,  $m$  and  $n$  so as to maximize  $U(y(1 - \beta_y), m) - c(n(1 + \tau(\bar{n})))$  subject to the budget constraint  $y + m = wn + \pi + M/P$ , where  $\pi$  is the household's profit income,  $M$  the supply of money, and  $P$  the price level, all of which the household takes as given. In equilibrium,  $m = M/P$  so that  $y$  and  $M/P$  must satisfy the marginal condition

$$(21) \quad U_1(y(1 - \beta_y), M/P)(1 - \beta_y) \\ = U_2(y(1 - \beta_y), M/P).$$

Given any  $(w, \bar{n})$  the household's supply of

labor is uniquely determined by the marginal condition

$$(22) \quad w = c'(n(1 + \tau(\bar{n}))) (1 + \tau(\bar{n})) / \lambda,$$

where  $\lambda$  is the value of  $U_2$  when (21) is satisfied. Homogeneity implies that  $\lambda$  is a constant. Equation (22) can be solved for the "effective" supply-of-labor function  $n^s(w, \bar{n})$ , with

$$(23) \quad n^s > 0, \quad n_1^s > 0 \quad n_2^s > 0.$$

The dependency of labor supply upon the realized demand for labor,  $\bar{n}$ , arises because an increase in  $\bar{n}$  reduces the per unit cost of selling labor. Since  $\bar{n}$  is inversely related to the rate of unemployment (16), therefore the dependency of labor supply upon  $\bar{n}$  is operationally the same as the "discouraged worker" effect.

An equilibrium is defined as a triple  $(w, n, y)$  such that

$$(24) \quad n^d(w, y) = n^s(w, n),$$

$$(25) \quad n = n^s(w, n),$$

$$(26) \quad y(1 + \sigma(y)) = f(n(1 - \beta_n)).$$

Equation (24) is the usual labor market equilibrium condition. But equilibrium requires also that the quantity signals to which the agents are responding correspond to the actual market quantities. Thus (25) requires the quantity of labor taken as given by the household to equal the equilibrium quantity, and (26) requires the sales quantity taken as given by firms to equal the equilibrium quantity; that is the amount produced  $f(n(1 - \beta_n))$  minus the amount used up in marketing  $y\sigma(y)$ . (I now omit bars from market quantities.) Given the equilibrium values of these variables, the price level satisfying (21) will equate  $y$  with the demand for output.

The equilibrium condition (26) can be rewritten in the following way. First define the function  $q(y) \equiv (1 + \sigma(y))y$ . This function indicates the amount of production required to market any given quantity of sales. By (9)

<sup>12</sup>In this sense the model is similar to the stochastic manipulable rationing model of Lars Svensson (1980), in which the cost to an agent of trying to exceed his allocated ration is that he might be forced to make the proposed trade. The difference is that, in Svensson's model, the rationing process uses up no resources whereas in this model it is costly to propose trades. In both cases one ends up with an alternative concept of effective demand.

and (11),

- (27) For all  $y > 0$ ,  $q(y)$  is continuous,  
strictly increasing, with  $q(y) > y$ .

Next, define  $\underline{q} \equiv \lim_{y \rightarrow 0} q(y) \geq 0$ , and  $\underline{n} \equiv (1 - \beta_n)^{-1} f^{-1}(\underline{q}) \geq 0$  (where  $f^{-1}(0) \equiv 0$ ). Then  $\underline{q}$  is the minimal amount of output needed to market any sales at all, and  $\underline{n}$  the corresponding amount of employment. Note that  $\underline{q}$  is well defined, by (27), that  $\underline{n}$  exists, by (17), and that  $\underline{n} = 0$  if and only if  $\underline{q} = 0$ , by (17).

Next I show that given  $n$ , (26) has a solution  $y > 0$  if and only if  $n > \underline{n}$ . To show this suppose first that  $n \leq \underline{n}$ . Then, by (17) and (27)  $f(n(1 - \beta_n)) \leq f(\underline{n}(1 - \beta_n)) \equiv \underline{q} < q(y)$  for all  $y > 0$ , so (26) has no such solution. Next suppose that  $n > \underline{n}$ . Then by (17) and (27):

$$\lim_{y \rightarrow 0} q(y) \equiv f(\underline{n}(1 - \beta_n)) < f(n(1 - \beta_n))$$

$$< \infty = \lim_{y \rightarrow \infty} q(y).$$

So, by Roll's theorem and the continuity of  $q(\cdot)$ , there is a solution to (26) with  $y > 0$ .

Finally note that, since  $q(\cdot)$  is strictly increasing, the implicit function theorem implies that (26) defines a function  $\bar{y}(n)$  for all  $n > \underline{n}$  and that

$$(28) \quad \bar{y}'(n) = (1 - \beta_n) f' / q'$$

$$= \frac{(1 - \beta_n) f'}{1 + \sigma + y \sigma'} > 0,$$

$$(29) \quad \lim_{n \rightarrow \underline{n}} \bar{y}(n) = \lim_{n \rightarrow \underline{n}} q^{-1}(f(n(1 - \beta_n)))$$

$$= \lim_{x \rightarrow \underline{q}} q^{-1}(x) = 0.$$

The function  $\bar{y}(\cdot)$  indicates the level of sales that can feasibly be marketed given any amount of employment greater than the minimal amount  $\underline{n}$ . Condition (26) is equivalent to the condition:  $y = \bar{y}(n)$ .

I can now reduce the definition of equilibrium to one involving a single equation in

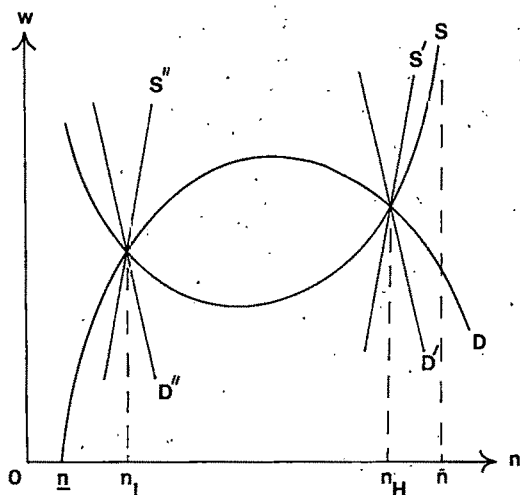


FIGURE 2

$n$ . From the definitions of  $n^d$  and  $n^s$ ,  $(w, n, y)$  is an equilibrium if and only if  $n$  satisfies

$$(30) \quad \frac{1 - \beta_n}{1 + \sigma(\bar{y}(n))} f'(n(1 - \beta_n))$$

$$= (1 + \tau(n)) c'(n(1 + \tau(n))) / \lambda,$$

$w$  equals the common value of either side of (30), and  $y = \bar{y}(n)$ . The two sides of (30) are the demand price and supply price for  $n$  units of labor, represented by  $D$  and  $S$  in Figure 2.

Because of the externalities, the equilibrium will not generally be unique. To see this suppose first that there were no externalities—that  $\sigma$  and  $\tau$  were constant. Then the usual labor market equilibrium condition (24) would be sufficient to determine the equilibrium level of employment, without the quantity equations (25) and (26), because  $n^s$  and  $n^d$  would depend only upon  $w$ . This equilibrium condition would be equivalent to (30) with  $\sigma$  and  $\tau$  constant. Equilibrium would be unique because the assumptions of declining marginal product (17) and rising marginal disutility (20) of labor would guarantee that the demand price was decreasing and the supply price increasing in  $n$ . This is illustrated by  $D'$  and  $S'$  in Figure 2, along which  $\sigma$  and  $\tau$  equal  $\sigma(\bar{y}(n_H))$  and

$\tau(n_H)$ . But, with the externality,  $D$  can be upward sloping because as employment increases, the unit cost of selling output  $\sigma(\bar{y}(n))$  decreases (from (9) and (28)), which tends to increase the demand price. Likewise  $S$  can be downward sloping because as employment increases, the unit cost of selling labor decreases (from (15)) which tends to lower the supply price. These effects of externalities make it possible for  $D$  and  $S$  to intersect more than once in nonpathological cases.

Indeed it is possible to specify boundary conditions that guarantee multiplicity except in razor's-edge cases. For example, suppose that in order to sell any output at all, firms must produce at least some positive critical mass of output; that is, that  $\underline{q} > 0$  and  $\underline{n} > 0$ . Under this assumption the unit cost of selling rises to infinity as sales fall to zero, and rises so fast that the total cost  $y\sigma(y)$  is bounded away from zero. The example (12) specified above satisfies this assumption.

In order for  $n$  to be the equilibrium level of employment, all the expressions in the equilibrium condition (30) must be well defined. This implies that  $n \geq \underline{n}$ . It also implies that  $n(1 + \tau(n))$  must be less than the upper limit on labor supply  $\bar{l}$ . From (15) the function  $l(n) \equiv n(1 + \tau(n))$  is strictly increasing and positive valued for all  $n > 0$ , with a range  $(l, \infty)$  for some  $l \geq 0$ . Assume that  $l < \bar{l}$ . Then  $\bar{n} \equiv l^{-1}(\bar{l})$  is well defined and the restriction imposed by the limit on labor supply is that  $n$  be less than  $\bar{n}$ .

If  $\underline{n} \geq \bar{n}$ , then no equilibrium is possible. So suppose that  $\underline{n} < \bar{n}$ . Then the equilibrium level of employment must lie in the interval  $(\underline{n}, \bar{n})$ ; there must be enough to get the output market started but no more than can feasibly be supplied. Therefore, to show generic multiplicity, it suffices to show that the supply price exceeds the demand price for values of  $n$  close enough to  $\underline{n}$  and for values close enough to  $\bar{n}$ . If this is so, then in terms of Figure 2 the only way that  $D$  and  $S$  can have a unique point of intersection is for them to be tangent at that point. But this is obviously a razor's-edge case. Generically there will exist an even number of intersections, so that if equilibrium exists it will be nonunique.

To show this, consider first what happens as  $n$  approaches  $\bar{n}$ . Because  $\bar{n} > \underline{n} > 0$ , the demand price remains bounded. But  $n(1 + \tau(n)) \equiv l(n)$  approaches  $l(\bar{n}) \equiv \bar{l}$  so that, by (20), the supply price becomes infinite. Next suppose  $n \rightarrow \underline{n}$ . Because  $\underline{n} > 0$ , (15) and (20) imply that the supply price remains bounded away from zero. Also, from (17),  $f(n(1 - \beta_n)) \rightarrow f(\bar{n}(1 - \beta_n)) > 0$ ; but from (29),  $\lim_{n \rightarrow \underline{n}} (1 + \sigma(\bar{y}(n))) = \lim_{y \rightarrow 0} (1 + \sigma(y)) = \infty$ , so the demand price approaches zero. Therefore the demand price is less than the supply price as  $n$  approaches either limit.

A similar demonstration of generic multiplicity goes through under the alternative assumption that  $\bar{l} \equiv \lim_{n \rightarrow 0} (1 + \tau(n))n > 0$ . Even with both  $\bar{l}$  and  $\underline{q}$  equal to zero a similar demonstration can be constructed under the assumption that

$$\lim_{n \rightarrow 0} [(1 + \tau(n))c'(n(1 - \tau(n))) \\ (1 - \sigma(\bar{y}(n)))] \\ / [\lambda(1 - \beta_n)f'(n(1 - \beta_n))] > 1.$$

Another consequence of the externalities is that not all gains from trade are fully exploited in an equilibrium even if it is unique. Specifically, suppose firms were to hire a small amount more labor at the same wage, and the price level were to adjust so as to allow the resulting increase in output to be sold. Consider what would happen to the typical firm's profit:  $f(n(1 - \beta_n))/(1 + \sigma(\bar{y}(n))) - wn$ . If there were no externalities (i.e., if  $\sigma$  were constant) then the firm's marginal condition (18) would imply that profit would remain unchanged. But, with the externalities,  $\sigma$  would fall, so profit would increase. Next, consider the level of household utility. If  $\sigma$ ,  $\tau$ , and  $m$  were constant, then according to (28)  $y$  would change at the rate  $dy/dn = (1 - \beta_n)f'/(1 + \sigma)$ , which by the firm's marginal condition equals the wage  $w$ ; the change in household utility would therefore be  $(1 - \beta_n)U_y w - (1 + \tau)c'$ , which, by the marginal condition (22) and the definition of  $\lambda$ , equals zero. But, with the externality,  $\sigma$  and  $\tau$  would decrease. The decrease in  $\sigma$  would make  $y$  increase by more than  $w$ . The decrease in  $\tau$  would make the disutility of

labor rise by less than  $(1 + \tau)c'$ . Furthermore, the change in the price level would have to keep the marginal condition (21) satisfied, which by homogeneity would require real balances  $m$  to equal  $k\bar{y}(n)$  for some constant  $k$ . Therefore  $m$  would increase. Thus taking into account the externalities and the change in  $m$ , utility would increase. Both firms and households would gain.

The following argument shows that, if there are multiple equilibria, household utility will be strictly greater whenever employment is greater. Consider any two equilibria,  $n_0$  and  $n_1$ , each satisfying (30), with  $0 < n_0 < n_1$ . Define

$$\phi(n) \equiv U(\bar{y}(n)(1 - \beta_y), k\bar{y}(n)) - c(n(1 + \tau(n))).$$

I want to show that  $\phi(n_0) < \phi(n_1)$ . Next, define

$$\Psi(n) \equiv \max_{\{x\}} U\left(f(x(1 - \beta_n)) \frac{1 - \beta_y}{1 + \tau(\bar{y}(n))}, k\bar{y}(n)\right) - c(x(1 + \tau(n))).$$

The first-order condition uniquely defining the solution to this problem is

$$U_1\left(f(x(1 - \beta_n)) \frac{1 - \beta_y}{1 + \sigma(\bar{y}(n))}, k\bar{y}(n)\right) \cdot f'(x(1 - \beta_n)) \frac{(1 - \beta_y)(1 - \beta_n)}{1 + \sigma(\bar{y}(n))} - c'(x(1 + \tau(n)))(1 + \tau(n)) = 0,$$

which is satisfied by  $x = n$  whenever  $n$  satisfies (30). Thus  $\Psi(n) = \phi(n)$  whenever  $n$  satisfies (30). Furthermore, for all  $n > 0$ ,

$$\Psi'(n) = -\left[\sigma' \cdot \bar{y}' \cdot f \cdot (1 - \beta_y) / (1 + \sigma)^2\right] U_1 + k\bar{y}' U_2 - \tau' \cdot x \cdot c' > 0.$$

Therefore,  $\phi(n_0) = \Psi(n_0) < \Psi(n_1) = \phi(n_1)$ .

By (16) the rate of unemployment will be lower whenever employment is greater. But what happens to the total amount of unemployment  $n\tau(n)$  as employment increases cannot be predicted.

#### IV. Comparison to the Keynesian Account

The preceding example shows how transaction costs and the externality of thin markets might account for persistently high rates of unemployment. Under any of the boundary conditions discussed, or more generally whenever multiple equilibria exist, the economy could shift permanently from one equilibrium to another with a lower level of employment and a higher rate of unemployment. How such shifts might be initiated and what course they would take can't be analyzed with this static model. But the model predicts that increases in unemployment can persist indefinitely.

The example has several Keynesian features, all of which are attributable to the fact that each agent is affected directly by the quantities  $\bar{n}$  and  $\bar{y}$  chosen by others. First, as we have seen, these direct effects are the reason why the example exhibits Keynesian "effective" demand and supply functions for labor.

Second, in the case of multiple equilibria the model exhibits a reciprocal feedback between labor and output markets similar to that involved in the Keynesian multiplier process. Consider a shift from  $n_H$  to  $n_L$  in Figure 2. This can be described as leftward shifts in the "constant-selling-cost" demand and supply functions to  $D''$  and  $S''$ . The level of aggregate demand declines because of the decline in employment (from  $\bar{y}(n_H)$  to  $\bar{y}(n_L)$ ). Conversely, employment declines at least in part because of the leftward shift in labor demand, which is caused by the decline in aggregate demand  $\bar{y}$ .

The case of multiple equilibria highlights another Keynesian feature;<sup>13</sup> namely, that the example is consistent with no discernible

<sup>13</sup>"Keynesian" in the sense of the account described in Figure 1. Keynes' *General Theory* actually implies countercyclical wages.

cyclical pattern to real wages. It is obviously possible for the real wage to be the same in the two equilibria of Figure 2, in which case employment can fluctuate (between  $n_L$  and  $n_H$ ) with no change in the real wage.

Another Keynesian feature is the result that not all gains from trade will be fully exploited in equilibrium. We saw that, as in the Keynesian account, firms would willingly hire more workers if they thought that aggregate demand would go up by as much as the net supply of output; but they have no incentive to do so because they take the level of aggregate demand as given.

Finally, it can be argued that the unemployment exhibited by the model is involuntary, as in Keynesian economics, despite the flexibility of wages. As Don Patinkin (1965, pp. 313–15) has argued, “involuntary” must mean chosen subject to an “unusually severe” constraint. By this criterion the unemployment in our example is involuntary in two distinct senses. First, in addition to the “usual” budget constraint workers are subject to a constraint requiring so much time to be spent in unemployment for every hour worked. Although each household is choosing to incur the amount  $n\tau(\bar{n})$  of employment, this choice is not being made as a substitute for employment as it would be if there were only the budget constraint. In this sense all unemployment in the example is involuntary.

In principle there is, however, no reason to think of the constraints imposed by transaction costs as being any less “usual” than the budget constraint. Still there is a sense in which at least some of the unemployment in a low-level equilibrium ( $n_L$  in Figure 2) fits Patinkin’s criterion of involuntariness. For, in this equilibrium, the transaction-cost constraint could be regarded as “unusually severe.” In order to sell  $n_L$  units of labor, the household is required to spend the amount  $n_L\tau(n_L)$  in unemployment instead of the smaller amount  $n_L\tau(n_H)$  that would be required to sell this much if the economy were at its high-level equilibrium. Thus the amount  $n_L(\tau(n_L) - \tau(n_H))$  might be regraded as involuntary. If we interpret these increased costs along the lines suggested by Okun’s analysis as the increased difficulty of

finding potential employers who are actively hiring, then they are indeed what observers of labor markets in depression refer to when describing the rise in unemployment as involuntary.

The example thus shows that it is possible to develop an account of unemployment with several Keynesian features, based upon the externalities of transaction costs, without some of the shortcomings of the Keynesian account. Specifically, the example does not require an assumption of unrealistically pessimistic sales conjectures as a foundation for its concept of effective demand, and it avoids the dilemma of basing a theory of “involuntary” unemployment on the refusal of the unemployed to work for less than the going wage.

This is not to say that the example provides a restatement of Keynesian economics. Among the Keynesian elements missing from it is any role of the marginal propensity to consume in the multiplier process. Nor is the model as it stands able to explain how changes in the level of aggregate demand, whether exogenous or induced by monetary or fiscal policy, could affect the level of output. In particular, money is neutral in the model, in the sense that all equilibria remain invariant in real terms to changes in  $M$ . Such changes might propel the economy from one real equilibrium to another, but the model as it stands has no dynamics with which to address the question. Nor is the model consistent with the Keynesian implication that, within some interval, any level of employment is consistent with equilibrium under the same tastes and technology. Furthermore the assumption of perfectly flexible wages and prices makes it inconsistent with Keynesian explanations of the slow adjustment of nominal wages and prices.

## V. Conclusions

Quantity adjustments are often seen as substitutes for price adjustments in Keynesian economics. If prices were perfectly flexible transactors would not have to be informed of realized quantities in order to formulate mutually consistent trading plans; they would just need to observe prices that

had been adjusted to their equilibrium values. If prices were inflexible transactors would also have to know realized quantities, which would have to be adjusted to their fixed-price equilibrium values. Patinkin (1976, especially pp. 65–66) has identified the essential contribution of Keynes' *General Theory* as the analysis of the equilibrating role of such quantity adjustments.

The example presented in this paper suggests that if there are significant external economies operating through transaction costs, quantity and price adjustments may be complements, not substitutes. Even with perfectly flexible prices, transactors must know the equilibrium values of realized quantities in order to formulate mutually consistent plans. The quantity signals to which they respond must satisfy the consistency conditions (25) and (26). Because many of the essential features of the Keynesian account of persistent unemployment follow from this need for quantity adjustment, they will be exhibited even in models with fully flexible prices.

None of this implies that perfect price flexibility is a good assumption, or that the issue of price flexibility has little quantitative importance for understanding fluctuations in unemployment. But it suggests that the issue of price flexibility is not crucial to many Keynesian results. When the transaction costs that are implicit in Keynesian analysis are made explicit, these results go through regardless of the degree of price flexibility. It also suggests, therefore, that from a Keynesian perspective a deeper understanding of unemployment will come from paying more attention to these transaction costs rather than attaching to "sticky prices" the blame for all communication problems in the economy.<sup>14</sup>

Future research along these lines will obviously require a more explicit treatment of price formation, along with a more explicit account of the institutional arrangements underlying the cost of transacting.

<sup>14</sup> Joseph Ostroy (1973) gives a graphic description of the communication problems that would remain even if an auctioneer costlessly computed and announced equilibrium prices.

## REFERENCES

- Azariadis, Costas, "Implicit Contracts and Underemployment Equilibria," *Journal of Political Economy*, December 1975, 83, 1183–202.
- Barro, Robert J., "Long-Term Contracting, Sticky Prices, and Monetary Policy," *Journal of Monetary Economics*, July 1977, 3, 305–16.
- and Grossman, Herschel I., "A General Disequilibrium Model of Income and Employment," *American Economic Review*, March 1971, 61, 82–93.
- Baumol, William J., "The Transactions Demand for Cash: An Inventory Theoretic Approach," *Quarterly Journal of Economics*, November 1952, 66, 545–56.
- Clower, Robert W., "The Keynesian Counter-revolution: A Theoretical Appraisal," in Frank H. Hahn and Frank P. R. Brechling, eds., *The Theory of Interest Rates*, London: Macmillan, 1965.
- Diamond, Peter A., "Aggregate Demand Management in Search Equilibrium," *Journal of Political Economy*, October 1982, 90, 881–94.
- Fischer, Stanley, "Long-Term Contracts, Rational Expectations and the Optimal Money-Supply Rate," *Journal of Political Economy*, February 1977, 85, 191–206.
- Hahn, Frank H., "Equilibrium with Transaction Costs," *Econometrica*, May 1971, 39, 417–39.
- , "On Non-Walrasian Equilibria," *Review of Economic Studies*, February 1978, 45, 117.
- Howitt, Peter W. and McAfee, R. Preston, "Search, Recruiting, and the Indeterminacy of the Natural Rate of Unemployment," Department of Economics, Research Report No. 8325, University of Western Ontario, December 1983.
- Iwai, Katsuhito, *Disequilibrium Dynamics*, New Haven: Yale University Press, 1981.
- Jones, Robert A., "The Origin and Development of Media of Exchange," *Journal of Political Economy*, August 1976, 84, 757–75.
- Laidler, David, "On Say's Law, Money, and the Business Cycle," in his *Monetarist Perspectives*, Oxford: Philip Allan, 1982.

- Leijonhufvud, Axel, *On Keynesian Economics and the Economics of Keynes*, New York: Oxford University Press, 1968.
- McCallum, Bennett T., "Price Level Adjustments and the Rational Expectations Approach to Macroeconomics Stabilization Policy," *Journal of Money, Credit, and Banking*, November 1978, 10, 418-36.
- Mortenson, Dale T., "Property Rights and Efficiency in Mating, Racing, and Related Games," *American Economic Review*, December 1982, 72, 968-79.
- Niehans, Jürg, "Money and Barter in General Equilibrium with Transactions Costs," *American Economic Review*, December 1971, 61, 773-83.
- Okun, Arthur M., *Prices and Quantities*, Washington: The Brookings Institution, 1981.
- Ostroy, Joseph M., "The Informational Efficiency of Monetary Exchange," *American Economic Review*, September 1973, 63, 597-610.
- Patinkin, Don, *Money, Interest, and Prices*, 2d ed., New York: Harper and Row, 1965.
- , *Keynes' Monetary Thought*, Durham: Duke University Press, 1976.
- Stiglitz, Joseph E., "Equilibrium in Product Markets with Imperfect Information," *American Economic Review Proceedings*, May 1979, 69, 339-45.
- Svensson, Lars E. O., "Effective Demand and Stochastic Rationing," *Review of Economic Studies*, January 1980, 47, 339-55.
- Woglom, Geoffrey, "Underemployment Equilibrium with Rational Expectations," *Quarterly Journal of Economics*, February 1982, 97, 89-107.



# The Demand for Unobservable and Other Nonpositional Goods

By ROBERT H. FRANK\*

The importance of demonstration effects in consumption behavior has long been recognized by economists and other social scientists.<sup>1</sup> But such demonstration effects by their very nature cannot apply with equal force to all categories of goods. We may know very well, for example, what kinds of cars acquaintances drive or the types of homes they live in, but we are much less likely to know how much they save or the amounts they spend on insurance.

Even in circumstances where what others consume is known, interpersonal comparisons with respect to certain *types* of consumption will be more important than will others. As Thorstein Veblen emphasized in 1899, at least some people appear actively concerned about how the amount of leisure they consume compares with the amounts consumed by their peers. But for most people, we may safely assume that such comparisons pale in relation to the corresponding comparisons regarding, say, the education of their children.

Following Fred Hirsch (1976, ch. 3), I use the term "positional goods" here to mean those things whose value depends relatively strongly on how they compare with things owned by others. Goods that depend relatively less strongly on such comparisons will be called nonpositional goods. As noted, the nonpositional category includes, but is not limited to, goods that are not readily observed by outsiders. This paper explores how patterns of spending behavior are affected by the fact that interpersonal comparisons apply

with greater force to some goods than to others.

Section I begins with an example that illustrates why interpersonal comparisons are more important for some goods than for others. This example also illustrates, in a qualitative way, the conclusion that noncooperative consumption decisions result in an underconsumption of nonpositional goods.

Section II then describes a formal model in which rank effects produce downward distortions in individual demands for nonpositional goods. Under certain circumstances, collective restrictions on consumption behavior are shown to produce welfare improvements, even for fully rational consumers operating in structurally competitive environments. Budget shares for certain nonpositional goods are shown to vary systematically with income and with the access individuals have to mechanisms for implementing cooperative consumption agreements.

Section III further explores the adaptive significance of imitative behavior. Such behavior is shown to be individually adaptive, but collectively maladaptive, in the context of a signaling competition in which observable consumption goods help identify individuals of high ability.

Section IV summarizes a variety of empirical evidence that bears on the hypotheses put forward in Sections I and II. This evidence suggests that James Duesenberry's relative income hypothesis (1949) was abandoned prematurely by the economics profession. It also suggests an alternative interpretation of the economic role of the trade union.

The paper concludes by noting that forced savings programs, safety regulation, overtime laws, and various other regulations of the labor contract may be interpreted as devices for mitigating the consequences of competi-

\*Associate Professor of Economics, Cornell University, Ithaca, NY 14853. I thank Phil Cook, Larry Seidman, Bob Hutchens, and Dick Thaler for helpful discussions.

<sup>1</sup>For an extensive list of citations, see my 1985 study, chs. 2 and 7.

tions between workers for favored positions in the income hierarchy. The apparent strength of consumption externalities suggests that supply siders are barking up the wrong tree when they say that income and consumption taxes introduce distortions into important economic decisions. Rather, such taxes alleviate existing distortions in those same decisions.

### I. Individual Consumption Decisions when Relative Standing Matters

In my 1985 study (ch. 2), I have argued that useful insights into people's economic behavior are afforded by the view that the utility function (or what psychologists would call the structure of motivation) was shaped by the forces of natural selection. By this view, the human nervous system is hard-wired with a panoply of tastes and aversions that contribute (or once contributed) to the individual's reproductive fitness. Sugar tastes sweet to us, for example, because having had an affinity for ripened fruit once contributed significantly to our primate ancestors' capacities to survive and leave offspring.

A more general implication of this view is that an element of almost overriding importance in the structure of human motivation will be a taste for seeing to it that one's children are launched in life as successfully as possible. Now, how successful one's children will be in life depends much less on their skills and endowments in any absolute sense than on how these compare with the skills and endowments of others. Success in the labor market, for example, depends much less on the quality of instruction one receives, *per se*, than on how one's training compares with the training received by others.<sup>2</sup> Suppose we take as a working hypothesis that a parent's utility function is programmed with an instruction something like, "Feel bad whenever your children are less well provided for than are the children of your peers." What sorts of behavior would

such a utility function predict that would not be predicted by the utility functions that economists generally work with?

To pursue this question, consider an example in which two persons, *A* and *B*, are each faced with the choice of working in a clean mine or a dusty mine. Wages in the clean mine at \$200 a week are lower than those in the dusty mine by \$50, an amount that reflects the cost of maintaining a dust-free working environment. The lone adverse consequence of working in the dusty mine is that life expectancy is shortened by fifteen years.

If *A* is strongly concerned about where his children stand vis-à-vis *B*'s (with respect to education and various other advantages), and if *B* feels that same concern, then the payoff to each from working in a given mine will depend in a clear way on the mine chosen by the other. In choosing between the two mines, each must weigh not only his feelings about the value of extended longevity in the abstract, but also the fact that his choice will affect his ranking in the income hierarchy. Suppose the two rank the four possible outcomes in the way shown in Table 1.

The rankings in the upper-left and lower-right cells of Table 1 indicate that, in the absence of concerns about the relative standing, each would find it worthwhile to sacrifice \$50 a week in order to escape working in the dusty mine. But neither is willing to make that same exchange if in the process he loses ground in the income hierarchy. As the rankings are configured here, *A* and *B* confront a standard example of the prisoner's dilemma. The dominant strategy for each is to choose the dusty mine. Yet, by so doing, an outcome results that each finds distasteful in comparison with the (feasible) alternative of both working in the clean mine.

If preferences were indeed forged in the crucible of natural selection, it is easy to see why people might find it attractive to sacrifice longevity in return for an opportunity to provide decisive advantages for their children (or in order to prevent their children from becoming seriously disadvantaged in a relative sense). Yet the number of favored positions in any rank ordering is fixed inescapably by the laws of simple arithmetic. And

<sup>2</sup>See for example, A. Michael Spence (1974), Hirsch, and Lester Thurow (1975).

TABLE 1—MINE SAFETY CHOICES WHEN  
RELATIVE STANDING MATTERS

A	B	
	Clean Mine	Dusty Mine
Clean Mine	Second best for A Second best for B	Worst for A Best for B
Dusty Mine	Best for A Worst for B	Third best for A Third best for B

Note: Clean mine: \$200 a week; Dusty mine: \$250 a week.

thus the exchange that is so attractive from each individual's point of view has no similar allure when viewed from the perspective of the population as a whole.

A related distortion is present when individuals make decisions about how much leisure to consume. To the extent that extra income is valued not only for its own sake, but also for the *relative* advantages it affords, the option of working an additional hour will appear misleadingly attractive to individuals.<sup>3</sup> Conventional economic analysis shows the workweek that emerges in an atomistically competitive labor market to be Pareto optimal, but when relative standing is a primary concern, this result no longer holds. For the perceived individual payoffs from the sale of leisure will then add up to more than the realized aggregate payoff.

Such distortions as these need not be viewed as having arisen because people are concerned about relative standing per se. As Hirsch, Amartya Sen (1983), and others have emphasized, having high relative standing is *instrumental* to the realization of numerous legitimate human objectives.<sup>4</sup> Disdainful attitudes towards people's efforts to "keep up with the Joneses" should not be allowed to obscure the fact that concerns about relative standing are completely consistent with the rational pursuit of self-interest.

<sup>3</sup>Duesenberry makes a similar point. See also Richard Layard (1980) and Michael Boskin and Eytan Sheshinski (1978).

<sup>4</sup>See also my 1985 study.

## II. A Simple Model of the Demand for Nonpositional Goods

Though the characteristics of consumption goods clearly vary continuously along many different dimensions, it will be convenient for analytical purposes to think of goods as falling into one of two classes, positional goods and nonpositional goods. Let us assume an individual's utility is determined by how much of each type of good he has and how his consumption compares with the consumption of others. Interpersonal comparisons matter, by definition, only with respect to positional goods. Specifically, let us assume a population of individuals in which all have identical utility indexes,

$$U = U(x, y, R(x)),$$

where  $x$  = positional consumption level,  $y$  = nonpositional consumption level, and  $R(x)$  is a number between 0 and 1 indicating the percentile ranking of  $x$  in the population of  $x$  values. If  $f(x)$  represents the density function for  $x$  values and  $x_0$  is the smallest value taken by  $x$  in the relevant population, then an individual with  $x = x_1$  will have

$$(1) \quad R(x_1) = \int_{x_0}^{x_1} f(x) dx.$$

When individuals are spoken of below as making consumption decisions noncooperatively, this will mean that they make the Nash-Cournot assumption that their own spending behavior does not perceptibly alter the spending behavior of others. That is, noncooperative consumption demands are defined as those that emerge when individuals maximize utility taking the density  $f(x)$  as being externally fixed.

The first-order conditions for the utility maximization exercise here are

$$(2) \quad (U_1/U_2) + (U_3/U_2)R'(x) = P_x/P_y,$$

and

$$(3) \quad P_x x + P_y y = M,$$

where  $U_i$  is the partial derivative of  $U$  with respect to its  $i$ th argument,  $P_x$  and  $P_y$  are the prices of  $x$  and  $y$ , and where  $M$  is income, which is exogenously given for each individual. Equation (1) says that  $R'(x) = f(x)$ , so equation (2) may be rewritten as

$$(2') \quad (U_1/U_2) + (U_3/U_2)f(x) = P_x/P_y.$$

Against the equilibrium condition given in (2'), let us now contrast the solution that emerges when individuals maximize utility by acting cooperatively. First we must specify what purpose cooperative behavior is meant to achieve in this context. In equation (2') we see that when individuals act noncooperatively, each perceives that additional consumption  $x$  augments utility not only through its direct effect,  $U_1$ , but also through its indirect effect on the rank term,  $R(x)$ . Yet the assumption of identical utility indexes assures that once the noncooperatively determined equilibrium is reached, each individual's ultimate ranking in the positional goods hierarchy will be the same as his original ranking in the exogenously given income hierarchy. Viewed from the perspective of the collective, the second, indirect return to positional goods consumption is thus entirely spurious.

Let us assume, therefore, that the objective of the cooperating population is to eliminate the influence of this spurious return from individual consumption decisions. If  $g(m)$  represents the original density function of income values and  $m_0$  the smallest income level in the population at issue, a natural way of accomplishing this objective is for each individual to allocate his income  $M$  across  $x$  and  $y$  as he would if his rank in the positional goods hierarchy were taken to be fixed in advance at

$$(4) \quad R(x) = \int_{m_0}^M g(m) dm = G(M).$$

That is, let us assume that the cooperative case may be thought of in terms of a collection of individual maximization problems of

the form

$$(5) \quad \max_{x,y} U(x, y, G(M)),$$

$$\text{subject to} \quad P_x x + P_y y = M.$$

Equation (5) is, of course, the same as the simple utility maximization problem from the traditional independent preferences setting, and its first-order conditions are thus

$$(6) \quad U_1/U_2 = P_x/P_y,$$

with the same budget constraint as in equation (3).<sup>5</sup>

Comparing the equilibrium equations (2') and (6), the following propositions may be easily established:

**PROPOSITION 1:** *Cooperatively determined demands will be higher for nonpositional goods and lower for positional goods than the corresponding demands determined noncooperatively.*

**PROPOSITION 2:** *Each individual's utility level will be higher in the case of cooperatively determined demands than in the case of noncooperatively determined demands.*

Both of these propositions reflect the fact that the presence of  $R(x)$  in the utility function acts as an implicit subsidy to positional goods consumption in the noncooperative case, with the usual attendant consumption distortions and welfare reductions.

Using the notation  $E_{U_i}$  to represent the elasticity of  $U$  with respect to its  $i$ th argument, and  $E_{R_x}$  to represent the elasticity of  $R(x)$  with respect to  $x$ , equations (2') and (6) can be rewritten as

$$(2'') \quad (y/x)(E_{U_1}/E_{U_2} + (E_{U_3}/E_{U_2})E_{R_x}) = P_x/P_y,$$

<sup>5</sup> It is easily shown that the allocation that emerges in the cooperative case lies in the core.

and

$$(6') \quad (y/x)(E_{U1}/E_{U2}) = P_x/P_y.$$

In populations in which  $x_0$ , the smallest value of  $x$ , exceeds zero,  $E_{R_x}$  will be infinite at  $x_0$  and, for any  $f(x)$  likely to be observed in practice, decline monotonically to zero as  $x$  moves toward the maximum value in its domain. Let us suppose that  $E_{R_x}$  behaves in this fashion, and let  $y = \lambda(x)$  represent the income expansion path that obtains for the cooperative case. Then, for all utility functions for which  $(E_{U3}/E_{U2})E_{R_x}$  is a decreasing function along  $y = \lambda(x)$  (a very unrestrictive condition, since  $E_{R_x}$  declines monotonically from  $\infty$  to 0 along that path), we may easily demonstrate

**PROPOSITION 3:** *Budget shares for nonpositional goods grow more rapidly (or decline less rapidly) with income in the noncooperative than in the cooperative case.*

As a special case of Proposition 3, let us consider the nonpositional good "savings," and suppose that, except for the influence of rank effects, savings behavior would be governed by the forces contemplated in either the permanent income or life cycle hypotheses of savings. That is, holding  $R(x)$  fixed, suppose that  $U(x, y, R(x))$  is homothetic in  $x$  and  $y$ , where  $x$  is now consumption and  $y$  is savings. With  $R(x)$  fixed, budget shares devoted to savings will then be constant across income levels. This means that Proposition 3 can be restated as

**PROPOSITION 3':** *Noncooperative budget shares for savings are an increasing function of the individual's rank in the income hierarchy of the population of which he is a member.*

To go further in assessing the quantitative differences between noncooperative and cooperative demands, let us impose additional restrictions on the form of the utility index,  $U$ . The Cobb-Douglas form has the homothetic property assumed for the savings example, and is analytically convenient.

Specifically, let

$$(7) \quad U(x, y, R(x)) = x^{\alpha_1} y^{\alpha_2} (R(x))^{\alpha_3},$$

where  $\alpha_1, \alpha_2$ , and  $\alpha_3 > 0$ .

Using equation (7), the first-order conditions for a maximum in the noncooperative case are given by the budget constraint from equation (3) and by

$$(8) \quad \alpha_1 y / \alpha_2 x + \alpha_3 y f(x) / \alpha_2 R(x) = P_x / P_y.$$

For illustrative purposes, suppose the density  $f(x)$  is uniform on the interval  $[x_0, Kx_0]$ , where  $K$  is some positive integer. Using this form for  $f(x)$ , equation (8) becomes

$$(9) \quad \alpha_1 y / \alpha_2 x + \alpha_3 y / \alpha_2 (x - x_0) = P_x / P_y, \\ \text{for } x \in [x_0, Kx_0].$$

The corresponding first-order condition for the cooperative case is simply

$$(10) \quad \alpha_1 y / \alpha_2 x = P_x / P_y.$$

Whether the demand functions that emerge from the cooperative case differ substantially from the corresponding noncooperative demand functions is thus seen to depend critically on the magnitude of the parameter  $\alpha_3$ , the elasticity of utility with respect to rank in the positional goods hierarchy. For the particular case in which  $\alpha_1 = \alpha_2 = \alpha_3$ , budget shares for nonpositional goods are as depicted in Figure 1.

As indicated in Figure 1, the budget share for nonpositional goods approaches zero for individuals near the bottom of the positional goods hierarchy, even though the derivative  $\partial U / \partial y$  becomes infinite as  $y$  approaches zero in the Cobb-Douglas form. Though the payoff to consuming  $y$  is very high at small values of  $y$  here, the payoff to additional consumption of  $x$  is even higher because of the advancement it enables in the positional goods hierarchy. That the hoped-for advance does not materialize in the end because of the parallel actions of others makes this con-

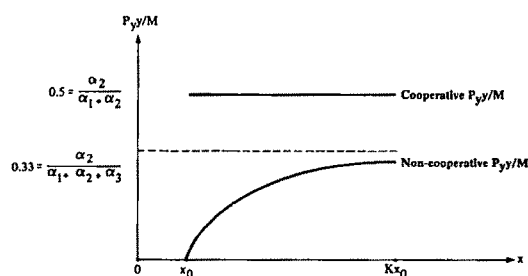


FIGURE 1. COOPERATIVE AND NONCOOPERATIVE BUDGET SHARES FOR UNOBSERVABLES

sumption behavior no less purposeful from the perspective of the individual; failure to allocate consumption in this fashion while others do would result in a backward movement in the positional goods hierarchy, hardly a better result under the assumed preference ordering.<sup>6</sup>

Elsewhere I have surveyed available empirical evidence from a number of studies on the comparative contributions to individual happiness levels made by absolute income levels on the one hand, and relative standing

on the other.<sup>7</sup> All of the findings reported in these studies are consistent with the hypothesis that relative standing is far more important than the absolute level of consumption in determining individual well-being. In view of the breadth and consistency of the available evidence on this question, it is far from fanciful to assign a significant role to the relative standing parameter  $\alpha_3$  in equation (7). If the elasticity of utility with respect to relative standing in the consumption hierarchy is considerably greater than that for absolute levels of goods consumption, as all available evidence suggests, the resultant consumption distortions, and their implied welfare consequences, will be even larger than those pictured in Figure 1.

### III. Consumption as a Signal of Ability

In the struggle to see the next generation safely launched, imitative behavior may, as noted in Section I, have individually adaptive consequences. Granted, by spending more on one's child's education today, one's rank in the consumption hierarchy may decline during retirement years. But from each individual's perspective, the decline in future ranking may be more than compensated for by the present gain. (The fact that the sought-after advance in the current rankings cannot be realized collectively is cold comfort to the individual who fails to keep rank today. I have argued elsewhere, 1985, ch. 7, that the divergence between individual and collective payoffs here may help account for what Pigou called the faulty telescopic faculty.)

From the individual's perspective, it does not even follow that consuming more now will necessarily result in diminished future consumption, because the information implicit in present consumption levels may affect future incomes. In societies in which

<sup>6</sup>If people are certain of their rank in the positional goods hierarchy, the model as it is expressed above does not produce a stable outcome. The lowest-ranking member of the hierarchy could initially move past the second lowest-ranking member by increasing his consumption of positional goods; and the second lowest-ranking member could then restore the original ordering by carrying out a similar shift of his own. But then the lowest-ranking member could reduce his consumption of positional goods without adversely affecting his ranking, which would already be as low as it could get. In turn, the second lowest-ranking member could then reduce his consumption of nonpositional goods without penalty, and in like fashion the higher-ranking members would one-by-one have an incentive to follow suit. Their having done so, the cycle would be set to begin anew.

Alternatively, the lowest-ranked member of the hierarchy may not be certain he is lowest ranked, and thus may be reluctant to act as if spending less on positional goods will not adversely affect his ranking. Another simple modification to the model that would generate a stable equilibrium would be to add the plausible assumption that people care not only about their rankings, but also about where they stand vis-à-vis the mean or some other cardinal parameter of the distribution. These modifications complicate the exposition, but do not alter the conclusions stated in Propositions 1-3 above.

<sup>7</sup>See my 1985 study, ch. 2. I have also argued that an implicit market exists for high-ranked positions in the earnings hierarchies we call firms. For all of the specific occupations for which I was able to construct empirical estimates, the implicit price of such positions is a substantial fraction of total earnings (see my 1984 article).

economic and social interactions between individuals are important (which is to say, in every society), information about the others with whom one might interact has obvious value. The mates we choose, the employees we hire, the people whose company we seek—all depend in a clear way on the information we are able to gather about other individuals in our environment. Many of the most important decisions ever made about us turn on others' estimates of what talents, abilities, and other characteristics we possess. Consider a population in which individuals' abilities are known to differ substantially but in which any specific individual's ability cannot be observed directly. Even in a loosely competitive labor market, there will be a strong positive correlation between individual ability and income levels. Similarly, when there is broad dispersion in income levels, there will generally be a strong positive correlation between individual income levels and various observable consumption goods: the size and location of one's home, the quality of one's automobile or wardrobe, the clubs to which one belongs, and so on. When an individual's ability level cannot be observed directly, such observable components of his consumption bundle constitute a signal to others about his total income level, and on average, therefore, about his level of ability.<sup>8</sup>

Let us explore the extent to which imperfect information about ability might create incentives for people to rearrange consumption patterns to favor observable goods. Consider a population of  $N$  individuals with productive ability levels  $A_1, \dots, A_N$ , which cannot be observed directly. The individuals are hired by firms in competitive labor markets, and are paid money wages,  $M_1, \dots, M_N$ , that are based on the firms' estimates of their ability levels. Suppose, in particular, that

$$(11) \quad M_i = z\hat{A}_i + (1-z)A_i, \quad i=1, \dots, N,$$

where  $0 < z < 1$ , and  $\hat{A}_i$  is the best estimate (in a sense to be defined presently) of indi-

vidual  $i$ 's ability that is available to persons outside the firm. That is to say, let us suppose that the wage a worker ultimately receives from the firm he works for is an unbiased amalgam of his true marginal product,  $A_i$ , and the best estimate thereof that was available to the firm when the worker was a job applicant. How job applicants look on paper may be of interest in its own right to employers (especially for jobs in which contact with people outside the firm is important), or may influence the extent to which firms invest in subsequent training for applicants.

Suppose that consumption of observable goods  $x_i$  is related to income  $M_i$  according to

$$(12) \quad x_i = g(M_i) + \gamma_i, \quad i=1, \dots, N,$$

where  $\gamma_i$  is a random term with  $E(\gamma_i) = 0$  and  $\text{var}(\gamma_i) = \sigma_\gamma^2$ . Faced only with information on  $x_i$ , an outside observer who knows the parameters that characterize the deterministic component on the right-hand side of equation (12) then has available an unbiased estimate of  $M_i$ , in the form of

$$(13) \quad \hat{M}_i = g^{-1}(x_i),$$

where  $g^{-1}(\cdot)$  denotes the inverse of the function  $g$ . Writing  $g = \beta(M_i) \cdot M_i = \beta_i M_i$ , and noting from equation (11) that the expectation of  $M_i$  equals  $A_i$ , the outside observer thus has an unbiased estimator of  $A_i$ , conditional on  $x_i$ , call it

$$(14) \quad \hat{A}_i^1 = x_i / \beta_i.$$

Now suppose the outside observer also has some other independent information about  $A_i$ . In particular, suppose there is a test  $T_i$  that satisfies

$$(15) \quad T_i = A_i + \tau_i, \quad i=1, \dots, N,$$

where  $\tau_i$  is a random term with  $E(\tau_i) = 0$  and  $\text{var}(\tau_i) = \sigma_\tau^2$ , for all  $i$ . The information in this test and the information about  $A_i$  from equation (14) can then be melded to form a composite estimate of  $A_i$ . From the

<sup>8</sup>Spence (ch. 8) attributes to Richard Zeckhauser the idea that consumption may act as a signal of ability.

stochastic independence of  $\tau_i$  and  $\gamma_i$ , it follows that the weighted sum

$$(16) \quad \hat{A}_i = \frac{\beta_i^2(1-z)\sigma_\tau^2}{\beta_i^2(1-z)\sigma_\tau^2 + \sigma_\gamma^2} \frac{x_i}{\beta_i} + \frac{\sigma_\gamma^2}{\beta_i^2(1-z)\sigma_\tau^2 + \sigma_\gamma^2} T_i,$$

is the minimum variance unbiased estimator for  $A_i$  in the class of linear combinations of  $T_i$  and  $x_i/\beta_i$ .

Given the ability estimate in equation (16), any individual can increase outsiders' estimates of his ability by devoting more of his resources to the purchase of  $x$ , according to

$$(17) \quad \frac{d\hat{A}_i}{dx_i} = \frac{\beta_i(1-z)\sigma_\tau^2}{\beta_i^2(1-z)\sigma_\tau^2 + \sigma_\gamma^2}.$$

The strength of this effect increases with the budget share for observables,  $\beta_i$ , and with the test variance,  $\sigma_\tau^2$ , and is inversely related to  $\sigma_\gamma^2$  and  $z$ . Unless the budget share for observables is very small, or the independent ability test extremely accurate, any one individual may substantially enhance others' estimates of his ability by increasing the share of his budget devoted to observables. For the particular case of  $z=1/2$ ,  $\beta_i=.8$ , and  $\sigma_\tau^2=\sigma_\gamma^2$ , the elasticity of  $\hat{A}_i$  with respect to  $x_i$  is more than .24, a very substantial effect indeed. Even when the effect on ability estimates of increasing  $x$  is much smaller than in the above example, it may nonetheless be sufficient to alter the outcome of important decisions regarding closely ranked candidates. Close employment decisions, for example, can obviously be influenced decisively even by very weak correlates of ability: placement counselors have long stressed the importance of quality attire and a good address in the job-search process.

To the extent that important outcomes do indeed hinge on the signals implicit in observable consumption levels, individuals who do not rearrange their consumption bundles in favor of observable goods will not always fare better than those who do. It may even

be the case that curtailing the *proportion* of income devoted to unobservable consumption goods will enhance an individual's earnings to such a degree as to raise the actual *level* of consumption of unobservables. Thus, while reduced consumption in the current period is normally thought of as enhancing consumption possibilities in later periods, precisely the opposite result may obtain if current consumption is an important indicator of ability. First impressions often count for a lot, and as the apparel companies are fond of reminding us, one doesn't get a second chance to make a first impression.

But while devoting extra resources to the consumption of observables may be highly adaptive from the point of view of the individual, it is clearly suboptimal from the point of view of the population as a whole. One individual's forward move in any hierarchy can occur only at the expense of backward moves by others. If some individuals rearrange their consumption bundles to favor observable goods, others who do not do so will then be perceived as standing lower in the distribution of productive ability than they actually do. One individual's "offensive" signal is cancelled by another's "defensive" signal, and in the end too many resources are devoted to the consumption of observable goods.

The ability-signaling rationale for imitative behavior suggests that incentives to distort consumption in favor of observable goods will be inversely related to the amount and reliability of independent information that exists concerning individual abilities. Stable environments in which long-standing social networks exist will have more such information than do less-stable environments, and for this reason the budget shares of unobservables should be larger in the former than in the latter. In the same vein, people who move frequently should have lower budget shares for unobservables than those who stay put.<sup>9</sup> To the extent that in-

<sup>9</sup>These observations are in accord with, and suggest a possible basis for, the fact that consumption patterns in small towns are often said to exhibit a certain sanity that metropolitan consumption patterns seem to lack.



dependent measures of an individual's ability become more numerous and reliable as an individual grows older, we expect budget shares for observables to decline with age. To the extent that individuals are in competition with one another for potential mates, budget shares for unobservables should be higher for married persons than for unmarried persons.<sup>10</sup>

Whether these predictions of the ability-signaling model will find empirical support remains to be seen. But even if imitative behavior could not be easily rationalized on the basis of the characterization of individual self-interest offered here, there would remain the question of whether the predictions about spending behavior made in Sections I and II are empirically valid. To this question let us now turn.

#### IV. A Survey of Empirical Evidence

##### A. *Savings vs. Income*

For many years economists struggled to resolve the apparent paradox implicit in the observation that the average propensity to consume falls with income in cross-section data, but is constant in time-series data. Duesenberry's proposed solution to this puzzle in 1949 was essentially the same as the one stated in Proposition 3' above, namely, that demonstration effects weigh relatively

more heavily on people with lower incomes, causing them to consume higher fractions of their incomes than do people with higher incomes. This sense of relative deprivation is not attenuated by across-the-board changes in absolute income, and Duesenberry thus saw no reason for aggregate income growth to alter the average propensity to consume over time.

Though Duesenberry's explanation was persuasive to many, and seemed an intuitively plausible description of how people actually behave, it is fair to say that many economists felt uncomfortable with what they regarded as a sociological theory of the consumption function. To many economists, the notion of consumers being strongly influenced by demonstration effects in consumption must have seemed troublingly at odds with the postulate of rational pursuit of self-interest. It is hardly surprising, therefore, that the profession later so warmly embraced Milton Friedman's permanent income hypothesis (1957) and the life cycle hypothesis of Franco Modigliani and Richard Brumberg (1955). Without relying on vague constructs borrowed from other branches of the social sciences, these theories provided clear a priori reasons, carefully grounded in utility-maximizing behavior, for the observed pattern of average propensities to consume in time-series and in cross-section data.

There is no question that the phenomena addressed by the permanent income and life cycle theories are real and important. But these theories simply cannot account fully for the positive relationship between savings rates and incomes we observe in cross-section samples of individuals. The life cycle and permanent income theories of saving both insist that if the influence of life cycle differences and transitory earnings could be eliminated, we would then see that high-income persons save the same fractions of their incomes as do low-income persons. In study after careful study, however, this prediction has failed to find empirical support. Thomas Mayer (1966), for example, has argued that one way of eliminating the effects of transitory earnings variations is to look at average savings rates across occupations. Though in any given year, for example, some

---

Differences between urban and rural consumption patterns may thus spring less from fundamental differences in the personal values held by the two groups than from differences in the payoffs they face from consuming observable goods.

<sup>10</sup> The importance of sending ability signals via the goods one consumes will naturally vary with one's chosen occupation. Earnings and the abilities that count most among research professors are not very strongly correlated, and many professors think nothing of continuing to drive a 10-year-old automobile if it still serves them reliably. But only in a very small town, where people know one another very well, might it not be a mistake for an aspiring young attorney to drive such a car in the presence of his potential clients. Good lawyers generally earn a lot of money, and people with a lot of money generally drive fashionable new cars. The potential client who doesn't know better is likely to assume that a lawyer with a battered car is not much sought after.

attorneys will have higher incomes than normal, others will have unusually low incomes; in a large sample of attorneys, therefore, the surpluses of those who had good years will largely cancel the shortfalls of those who had bad years. Mayer observed that the permanent income hypothesis requires that the average savings rate for an occupation should thus be independent of its average income level. He then gathered data on average savings rates and average income levels for different occupations in numerous Western countries during different periods in the twentieth century. For virtually every country for which the necessary data were available, Mayer found occupational savings rates positively correlated with average income levels by occupation, a pattern that is flatly inconsistent with the permanent income hypothesis.<sup>11</sup>

H. W. Watts (1958) went a step further by studying the savings behavior of groups of individuals selected so as to represent similar heterogeneous cross sections of the population with respect to age. In so doing, Watts eliminated not only transitory earnings effects by focusing on group averages, but life cycle effects as well. He notes that it is clear from his findings that other factors besides income affect savings rates, but that it is equally clear that there is a significant positive relationship between savings rates and lifetime income.

Perhaps the most damaging evidence against the life cycle and permanent income theories has come in a recent study by Peter Diamond and J. A. Hausman (1982). Using data that record the spending and savings behavior of the same group of individuals

over a multiyear period, Diamond and Hausman find that, even after accounting for permanent income and life cycle effects, savings rates still rise substantially with incomes. They write

...[O]ur most important finding is the extent to which the savings to permanent income ratio rises with permanent income. Not only does the level of savings (wealth) rise with permanent income, but it does so in a sharply non-linear fashion.... [for permanent incomes below \$4770 per year, the savings-permanent income ratio rises by 3.3 percent for each extra \$1000 of permanent income;] beyond \$4700 it rises 5.7% for each extra \$1000 and beyond \$12,076 it rises by 14.2%. These results strongly confirm... that a simple linear relationship between savings and permanent income is not supported in our data. ... [pp. 36-37]

Numerous other authors have presented evidence that savings rates are positively related to life cycle and permanent income.<sup>12</sup> In his review of this evidence, Mayer wrote, "...of all the many tests which have been undertaken by friends of the [proportional savings rate] hypothesis, *not a single one supports it*... I therefore conclude that the proportionality hypothesis is definitely invalidated..." (1972, p. 348).

The evidence on the savings vs. income relationship is so strong and so consistent that it would appear difficult for proponents of the permanent income and life cycle theories to continue to insist that savings rates are unrelated to income. Yet these claims persist in most major undergraduate and graduate texts in macroeconomics.<sup>13</sup>

I have argued here that, in contrast to the permanent income and life cycle theories, a consumption theory that incorporates people's concerns about relative standing is able

<sup>11</sup>Some authors (see, for example, P. L. Menchik, 1979) have attempted to reconcile the life cycle and permanent income hypotheses to the savings rate data by arguing that the rich are motivated to bequeath larger shares of their lifetime wealth to their heirs than are the poor. Yet the aggregate ratio of bequests to national income has not risen hand in hand with income as a consumption theory based on absolute wealth would require. If, on the other hand, the bequest motive depends on relative wealth, then the permanent income and life cycle theories are almost indistinguishable from Duesenberry's relative income theory.

<sup>12</sup>For a thoughtful survey of these studies, see Mayer (1972).

<sup>13</sup>At least two leading macroeconomics texts (Thomas Sargent, 1979, and Robert Gordon, 1978) do not even mention the relative income hypothesis at all.

to account for the observed positive relationship between savings rates and income. Granted, the permanent income and life cycle theories have made an important contribution to our understanding of consumer behavior—long-run considerations *are* important to most consumers, and anyone who ignores that fact will make systematic errors when trying to predict consumer behavior. But in view of the empirical evidence, the extent to which these theories have supplanted Duesenberry's relative income hypothesis in modern textbooks seems yet another testament to the power of the *a priori* beliefs held by most economists. This outcome is not without irony, since we have seen that concerns about relative standing may well be fully compatible with the rational pursuit of self-interest, and therefore presumably not at all in conflict with economists' important prior beliefs. If this view wins acceptance, it suggests that greater attention be accorded to Duesenberry's explanation of the savings rate paradox, at least until some new empirical evidence is uncovered that proves it faulty.<sup>14</sup>

### B. *Union vs. Nonunion Compensation Packages*

To examine Proposition 1 requires that we uncover some source of variation in the extent to which individuals are able to form cooperative consumption agreements with other members of their personal reference groups (the "relevant population" noted in Section II). Both union and nonunion firms commonly facilitate collective consumption agreements regarding insurance, savings, and a variety of other fringe items. Several considerations suggest, however, that union

members are relatively better positioned to implement such agreements than are their nonunion counterparts. First, the average length of job tenure is much higher for union than for nonunion members,<sup>15</sup> which presumably will give rise to closer personal associations between coworkers in union firms than in nonunion firms. Accordingly, a union member's personal reference group should be more heavily composed of coworkers than should the nonunion worker's. Second, a similar tendency should emerge as a result of union firms being larger, on average, than nonunion firms. Third, the very existence of the union's administrative apparatus may facilitate an exchange of information between coworkers that enhances the likelihood of their being able to form agreements about how compensation should be allocated between various budget categories.<sup>16</sup> These considerations suggest that budget shares devoted to nonpositional goods should be higher for union members than for nonunion members.

The sociological literature on reference group theory stresses that an individual's personal reference group tends to consist disproportionately of others who are similar in terms of age, education, and various other background variables.<sup>17</sup> We also know that union members earn significantly higher wages than do nonunion workers with comparable job skills.<sup>18</sup> These observations together imply that a union member with a given income level will have higher income relative to the noncoworkers in his personal reference group than will a nonunion worker with the same nominal level. Referring to Proposition 3, this union-nonunion difference in rank *vis-à-vis* noncoworker refer-

<sup>14</sup>Robert Clower was thus, in my view, correct when he wrote that "...there seems to be no reason why the basic Duesenberry ideas should not be accepted as an integral part of the pure theory of consumer behavior" (1952, p. 178.) But he went on to say "...one gets the impression... that the interdependence postulate is comparatively innocuous as concerns established doctrines; but this remark may require considerable qualifications in the light of subsequent, and perhaps more sophisticated, inquiries" (p. 178). This inquiry is hardly a very sophisticated one, but it does suggest a number of such qualifications.

<sup>15</sup>Jacob Mincer (1983) finds, for example, that quit rates in the union sector are about one-half as large as in the nonunion sector for young men and about one-third as large for men over 30.

<sup>16</sup>See, for example, the arguments advanced by Albert Hirschman (1970).

<sup>17</sup>See, for example, Robert Merton and Alice Kitt (1950), Leon Festinger (1954), James Davis (1959), and Robin Williams (1975).

<sup>18</sup>Mincer (1983), for example, finds ability-adjusted union wage premiums of 6–14 percent for men under 30, and 4–12 percent for older men.

ence group members will act to reinforce the specific predictions about union-nonunion differences in the shares of total compensation devoted to nonpositional goods.

Richard Freeman (1981) has examined the effect of collective bargaining on the fringe share of the compensation package, and his findings are strongly in accord with Proposition 1. Using data from the Bureau of Labor Statistics' *Expenditures for Employee Compensation Survey*, Freeman estimated the effect of collective bargaining on eight components of voluntary fringe benefits. These results are reproduced here as Table 2. The coefficients reported therein represent partial effects of unionism on the various fringe items, wage income having been included as an explanatory variable in the regression equation from which those coefficients were taken. Given what intuition tells us about what constitutes a positional consumption good, the coefficients in Table 2 are strikingly consistent with the hypothesized effect of unionism on the structure of compensation.

Note, for example, that collective bargaining has its largest impact on fringe items 1 and 4. Union workers devote almost 48 percent more to insurance benefits than do nonunion workers with the same income levels. Similarly, union workers devote more than 41 percent more to pensions than do nonunion workers with the same income levels. These findings are in strong accord with the hypothesis that cooperative decisions will tend to favor unobservable goods. The finding that union workers devote a larger share of total compensation to "paid" vacations than do similarly situated nonunion workers is consistent with the view that leisure is a nonpositional good.

Freeman's estimates of the effects of collective bargaining on shift differentials and overtime premiums offer a mixed message for the theory of collective bargaining offered here. That union members have higher shift differentials (for example, premiums for working at night) is consistent with the notion that union workers will act more effectively than others do to limit the extent to which they exchange unfavorable working conditions for higher incomes. Freeman re-

TABLE 2—THE EFFECT OF COLLECTIVE BARGAINING ON SPECIFIC FRINGES, ALL PRIVATE INDUSTRY, 1967–72

Fringe	Cents per Hour Spent on Fringe	
	(1)	Coefficients <sup>a</sup> (2)
1. Life, Accident, and Health Insurance	10.1	4.8 (0.2)
2. Vacation	8.3	1.6 (0.2)
3. Overtime Premiums	10.1	-0.5 (0.4)
4. Pension	9.4	3.9 (0.4)
5. Holidays	5.2	0.8 (0.1)
6. Shift Differential	1.1	0.3 (0.1)
7. Sick Leave	1.1	-0.5 (0.1)
8. Bonuses	1.8	-1.4 (0.3)

Source: Freeman (1981, Table 4, p. 503).

<sup>a</sup>For the effect of collective bargaining on col. 1. Standard errors are shown in parentheses.

ports, however, that overtime premiums are actually smaller for union workers than for nonunion workers. That finding does not support the view of union objectives offered here. But the union-nonunion difference is less than 5 percent of the total devoted to this fringe item, and is not statistically significantly different from zero. Overtime premiums, moreover, are largely dictated to employers by the provisions of the Fair Labor Standards Act, so it is not clear that we would expect to see significant union-nonunion differences in this item in any event. Sick leave is also smaller for union than for nonunion workers, though the difference here too is small.

Note, finally, that Bonuses (item 8) are substantially smaller for union workers than for nonunion workers with the same incomes. Bonuses are equivalent to wage income insofar as both come in the form of cash. Bonuses therefore represent a portion of the compensation package that is left free from any collective allocation pattern the respective groups may wish to promote. Accordingly, Proposition 1 predicts that the bonus item will be larger for nonunion than

for union workers who have similar income levels. And Freeman does find the former to be four and one-half times the latter.

Needless to say, other explanations than the one advanced here may be offered in support of the coefficient pattern we see in Table 2. Freeman's own explanation for the observed difference in fringe shares relies on the assumption that older workers are simultaneously less mobile and have greater demands for fringe benefits than do younger, less tenured workers. He asserts that the demands of more senior workers are effectively expressed through the collective bargaining mechanism, but tend to be understated in the competitive outcome, where the compensation package is shaped primarily by the preferences of younger, more mobile workers, who are relatively less concerned about fringe benefits. Let us briefly consider this alternative explanation.

Freeman's explanation requires that non-union employers, who are assumed to employ captive older workers, be unable to design a discriminatory compensation package that simultaneously appeals to the tastes of both junior and senior employees alike. If employers are free, as they appear to be, to offer compensation packages in which both wages and fringes can be linked by formula to the employee's length of tenure with the firm, then Freeman's nonunion firm must have higher labor costs than (and should eventually be driven out by) other nonunion firms that pay lower wages but provide greater fringe benefits to more senior workers. Indeed, union and nonunion establishments alike do in practice link both wage payments and at least some fringes, such as pensions and vacations, directly to length of tenure with the firm. Other fringes, such as life and accident insurance, are often linked to total compensation, which, in turn, is highly correlated with tenure. Just as Freeman's argument implies non-cost-minimizing behavior on the part of nonunion firms, it also requires non-utility-maximizing behavior on the part of unions. Since fringe packages are easily designed to discriminate by age, why should older union workers force younger workers to consume uneconomically large shares of compensation in the form of fringe benefits?

Mincer (1984) finds a pattern of union-nonunion compensation differences similar to the one found by Freeman, for which he offers yet another explanation. Mincer argues that union workers are fearful that if they raise wages too high, firms will find it profitable to constrain the number of hours employees may work. Mincer doesn't say, but the reason that unions don't simultaneously bargain with firms to prevent such hours reductions is perhaps that unpredictable variations in product demand (unobservable by workers) make it inefficient to do so. In any event, Mincer then argues that union workers try to frustrate this stratagem by demanding a larger share of their compensation in the form of fringe benefits, which act as lump sums in the compensation schedule, thus reducing the marginal gain to firms of curtailing hours worked.

This is a curious strategy for a union to pursue. Any union that had sufficient bargaining power to implement such a strategy presumably would also have sufficient power to demand and get a cash intercept term appended to its weekly salary formula. Shifting part of cash compensation into such an intercept term would produce the same change in marginal conditions facing firms as would shifting compensation into lumpy fringe benefits, but would afford workers greater latitude in their consumption decisions, and would lead therefore to higher utility levels for union workers.

Arguments similar to the ones discussed above may be applied to the comparison of safety levels across union and nonunion firms. Elsewhere (1985, ch. 7) I have argued that consumption decisions regarding "contingent goods" have many of the same properties as those that apply to nonpositional goods. A contingent good is one that has a payoff only if some unlikely event occurs. Insurance and safety devices are examples of such goods. If contingent goods are like nonpositional goods, then union workers should devote larger shares of total compensation to safety than should otherwise similar non-union workers.

Unfortunately, little reliable information exists on the total level of expenditures firms make to promote health and safety in the

workplace. But if union workers are better able to express cooperative demands for safety than are nonunion workers, it then follows that (holding income and the level of risk exposure constant) the reservation price for accepting a given increment in risk exposure should be higher for union than for nonunion workers.

In their widely cited 1976 paper, Richard Thaler and Sherwin Rosen report the results of a statistical study whose structure is well-suited for testing this hypothesis. In their study, they estimate that the union worker must receive a risk premium that is \$8.08 per week higher than the premium required by an identically situated nonunion worker for accepting a 1/1000 increase in the annual probability of death. The particular estimate of interest from the Thaler-Rosen study is their regression coefficient for the interactive effect of risk and union membership. A test of the null hypothesis that collective bargaining does not affect the wage-risk tradeoff translates in the Thaler-Rosen study as a test of the hypothesis that the coefficient on the (risk  $\times$  union) variable is zero. The *t*-statistic for this coefficient is 2.02, which enables us to reject the null hypothesis at conventional significance levels. The union-nonunion difference in risk premiums amounts to a substantial fraction (often more than one-half) of the total risk premium workers receive in return for the performance of risky tasks.<sup>19</sup>

W. Kip Viscusi (1980) has constructed an alternative explanation for higher union safety levels by arguing that union and nonunion workers have the same preferences over safety and wage income, but weigh the preferences of older, more risk-averse employees differently during the bargaining process. This explanation is identical in its structure to Freeman's explanation of why the fringe share of the compensation package is higher for union than for nonunion workers. And it suffers, therefore, from many of the same difficulties. In most firms there is a menu of different tasks to be performed,

and not all these tasks are equally risky. Under such circumstances, the preferences of risk-averse older workers can be accommodated by simply assigning younger workers to the relatively more risky tasks. That firms do not do so suggests that some factor other than age-related differences in preferences must explain the difference in safety levels between union and nonunion firms.<sup>20</sup>

The foregoing differences in the ways union and nonunion workers allocate their total compensation suggest an alternative interpretation of the role of the trade union movement. Many accounts of the trade union movement have stressed the role of unions as a force for neutralizing excessive market power in the hands of firms.<sup>21</sup> But if concerns about relative standing are as strong as they appear to be, the presence of monopsony power is not logically necessary to explain why individual workers might sell various aspects of their services too cheaply. When relative standing is important, there are sensible reasons, quite apart from the prospect of an increase in total compensation, why workers might seek to determine the distribution of compensation collectively rather than individually.

## V. Concluding Remarks

The interdependent choice framework discussed here suggests alternative interpretations of a variety of apparently paternalistic laws and regulations. The Social Security program, for example, has been defended on the grounds that consumers lack sufficient

<sup>19</sup>For a thoughtful survey of the literature on compensating wage differentials for exposure to risk, see Robert Smith (1979).

<sup>20</sup>Perhaps the older workers are behaving paternalistically (and altruistically, too, since it costs them money) toward the younger union workers. But altruism cannot account for the parallel implications of Viscusi's argument for behavior in nonunion firms. For why would nonunion firms inflict uneconomically large risk burdens on risk-averse older workers? In Viscusi's framework, both the firm and the older worker could do better by shifting the older workers from risky to less-risky tasks.

<sup>21</sup>For a completely unequivocal statement of this view, see John Mitchell (1903). More recent accounts paint a much broader picture of what trade unions do, but their role as a countervailing force to the market power of firms continues to be emphasized (as, for example, in the Viscusi and Freeman papers cited above).

foresight and self-discipline to save effectively for their retirement. But we have seen that such forced savings programs might have a coherent role to play even in a world populated by rigidly disciplined consumers with perfect foresight. The problem of inadequate savings arose here not because of character defects, but because of a divergence between individual and collective incentives to save.

Overtime laws, health and safety regulations, and a variety of other restrictions of competitive labor contracts have similarly been explained as devices needed to protect workers from being ravaged by avaricious monopsonists. The interdependent choice framework discussed here suggests the possibility of a useful role for those same institutions even in perfectly informed, atomistically competitive labor markets.

Now, it is easy to imagine the line of discussion pursued here being used to justify a host of egregiously meddlesome regulatory activities. Yet such a regulatory response would hardly be in keeping with the traditional remedies economists have proposed for problems that arise from the presence of externalities. If consumption externalities do indeed motivate many of the command-and-control regulatory interventions we currently observe, then a simple tax on positional consumption expenditures might attenuate the need for many of these interventions. If consumption externalities are as important as they appear to be, then supply sides have got matters turned completely around when they insist that income and consumption taxes introduce serious distortions into the labor-leisure choice. When relative standing is important, such taxes serve, on the contrary, to mitigate an already present distortion in that choice.

## REFERENCES

- Boskin, Michael J. and Sheshinski, Eytan, "Optimal Redistributive Taxation when Individual Welfare Depends upon Relative Income," *Quarterly Journal of Economics*, November 1978, 92, 589-600.
- Clower, Robert W., "Professor Duesenberry and Traditional Theory," *Review of Economic Studies*, No. 3, 1952, 19, 165-78.
- Davis, James A., "A Formal Interpretation of the Theory of Relative Deprivation," *Sociometry*, December 1959, 22, 280-96.
- Diamond, Peter A. and Hausman, J. A., "Individual Retirement and Savings Behavior," presented at SSRC-NBER Conference on Public Economics, Oxford, June 1982.
- Duesenberry, James S., *Income, Saving, and the Theory of Consumer Behavior*, Cambridge: Harvard University Press, 1949.
- Easterlin, Richard A., "Does Economic Growth Improve the Human Lot? Some Empirical Evidence," in Paul A. David and Melvin Reder, eds., *Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz*, Palo Alto: Stanford University Press, 1973, 89-125.
- Festinger, Leon, "A Theory of Social Comparison Processes," *Human Relations*, No. 2, 1954, 7, 117-40.
- Frank, Robert H., *Choosing the Right Pond: Human Behavior and the Quest for Status*, New York: Oxford University Press, 1985.
- \_\_\_\_\_, "Are Workers Paid their Marginal Products?," *American Economic Review*, September 1984, 74, 549-71.
- Freeman, Richard B., "The Effect of Unionism on Fringe Benefits," *Industrial and Labor Relations Review*, July 1981, 34, 489-509.
- Friedman, Milton, *A Theory of the Consumption Function*, NBER Conference Series, No. 8, Princeton: Princeton University Press, 1957.
- Gordon, Robert J., *Macroeconomics*, Boston: Little Brown & Co., 1978.
- Hirsch, Fred, *Social Limits to Growth*, Cambridge: Harvard University Press, 1976.
- Hirschman, Albert, *Exit, Voice and Loyalty*, Cambridge: Harvard University Press, 1970.
- Layard, Richard, "Human Satisfaction and Public Policy," *Economic Journal*, December 1980, 90, 737-49.
- Mayer, Thomas, *Permanent Income, Wealth and Consumption*, Berkeley: University of California Press, 1972.
- \_\_\_\_\_, "The Propensity to Consume Permanent Income," *American Economic Review*, December 1966, 56, 1158-77.
- Menchik, P. L., "Inter-generational Transmis-

- sion of Inequality," *Economica*, November 1979, 46, 349-62.
- Merton, Robert K. and Kitt, Alice S., "Contributions to the Theory of Reference Group Behavior," in R. K. Merton and P. F. Lazarsfeld, eds., *Continuities in Social Research. Studies in the Scope and Method of "The American Soldier"*, Glencoe: Free Press, 1950, 40-105.
- Mincer, Jacob, "Union Effects: Wages, Turnover, and Job Training," in R. G. Ehrenberg, ed., *Research in Labor Economics*, Suppl. 2, Greenwich: JAI Press, 1983, 217-52.
- , "The Economics of Wage Floors," in R. G. Ehrenberg, ed., *Research in Labor Economics*, 6, Greenwich: JAI Press, 1984, 311-34.
- Mitchell, John, *Organized Labor*, Philadelphia: American Book and Bible House, 1903.
- Modigliani, Franco and Brumberg, Richard, "Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data," in K. Kurihara, ed., *Post-Keynesian Economics*, London: Allen and Unwin, 1955, 388-436.
- Sargent, Thomas, *Macroeconomic Theory*, New York: Academic Press, April 1979.
- Sen, Amartya, "Poor, Relatively Speaking," *Oxford Economic Papers*, July 1983, 35, 153-69.
- Smith, Robert S., "Compensating Wage Differentials and Public Policy: A Review," *Industrial and Labor Relations Review*, April 1979, 32, 339-52.
- Spence, A. Michael, *Market Signaling: Informational Transfer in Hiring and Related Screening Processes*, Cambridge: Harvard University Press, 1974.
- Thaler, Richard and Rosen, Sherwin "The Value of Saving a Life: Evidence from the Labor Market," in Nestor Terleckyj, ed., *Household Production and Consumption*, NBER Studies in Income and Wealth, No. 40, New York: Columbia University Press, 1976, 265-98.
- Thurow, Lester, *Generating Inequality*, New York: Basic Books, 1975.
- van Praag, B. M. S., *Individual Welfare Functions and Consumer Behavior*, Amsterdam: North-Holland, 1968.
- Veblen, Thorstein, *The Theory of the Leisure Class*, New York: Macmillan, 1899.
- Viscusi, W. Kip, "Unions, Labor Market Structure, and the Welfare Implications of the Quality of Work," *Journal of Labor Research*, Spring 1980, 1, 175-92.
- Watts, H. W., "Long-Run Income, Expectations, and Consumer Savings," in *Studies in Household Economic Behavior*, Yale Studies in Economics, Vol. 9, New Haven: Yale University Press, 1958.
- Williams, Robin M., "Relative Deprivation," in Lewis A. Coser, ed., *The Ideal of Social Structures: Papers in Honor of Robert K. Merton*, New York: Harcourt Brace Jovanovich, 1975, 355-78.



# Fiscal Policy and Aggregate Demand

By DAVID ALAN ASCHAUER\*

This paper is an investigation of the effects of fiscal policy on private consumption and aggregate demand within an explicit intertemporal optimization framework. Empirical evidence is brought to bear on the following questions: Is consumption sensitive to the choice of tax versus debt financing of current government expenditure? To what extent, if any, does government spending directly substitute for private consumer expenditure?

The first question has stimulated a considerable amount of research since Robert Barro's (1974) revival of the "Ricardian equivalence" proposition. Lewis Kochin (1974), Barro (1978), J. Ernest Tanner (1979), and Roger Kormendi (1983) obtain empirical results favorable to the proposition that, to a first approximation, the choice between current taxation and debt issuance to finance a given government expenditure stream is irrelevant to the determination of the level of aggregate demand. On the other hand, Martin Feldstein rejects some of the assumptions adopted in the empirical specifications of Kochin, Barro, and Tanner, and comes to the conclusion that "... each of the basic implications of the so-called 'Ricardian equivalence theorem' is contradicted by the data" (1982, p. 9).

The second question has also been touched upon in recent empirical studies. Feldstein's (1982) results detract from the proposition of "fiscal neutrality" whereby an increase in government spending induces an *ex ante* crowding out of an equal amount of private consumption expenditure. However, Kormendi obtains support for his "consolidated approach" to fiscal policy by finding a substantial degree of substitutability between

government spending and private consumption.

The argument advanced in this paper is that probable misspecification bias in these previous studies renders the results suspect and may account for the fact that minor changes in the empirical models lead to radically different conclusions regarding the potency of fiscal policy. In place of the conventional methodology, an alternative approach is presented which exploits restrictions placed on the data by the first-order necessary conditions for intertemporal optimization in consumption. The empirical evidence is supportive of the joint hypothesis of rational expectations and Ricardian equivalence as well as of the proposition that government spending substitutes poorly for private consumption in utility.

## I. Effective Consumption and Intertemporal Optimization

Here I develop the model to be estimated later in the paper and point out some inherent difficulties in previous tests of fiscal neutrality. The theory applies to a representative individual who has time-separable preferences over private consumption,  $C$ , and the goods and services flowing from the government sector,  $G$ . Specifically, the agent's utility function is given by

$$(1) \quad V_t = \sum_{j=0}^{\infty} (1/(1+\delta))^j u(C_{t+j}^*),$$

where  $\delta$  is a constant rate of time preference and  $u(\cdot)$  is a time-invariant, concave momentary utility function. Finally,  $C_t^* = C_t + \theta G_t$  denotes the level of "effective" consumption in period  $t$ , a linear combination of private consumption and government goods and services. The constant marginal rate of substitution implies that a unit of govern-

\*Department of Economics, University of Michigan, Ann Arbor, MI 48109. I thank Robert King and members of the University of Michigan Money Seminar for useful comments. Any errors are my responsibility.

ment goods and services yields the same utility as  $\theta$  units of private consumption.<sup>1</sup>

The representative agent is allowed unrestricted access to a capital market at which he may accumulate or decumulate assets at the assumed constant real rate of interest  $r$ . His period  $t$  flow budget constraint is given by

$$(2) \quad W_{t+1}/(1+r) - W_t + C_t = N_t - T_t$$

where  $W_t$  = beginning of period holdings of one period bonds (which includes government debt), each unit of which is a claim to a unit of output,  $N_t$  = period  $t$  labor earnings and  $T_t$  = period  $t$  tax payments (net of transfers). Forward substitution in equation (2) yields

$$(3) \quad \sum_{j=0}^{\infty} (1/(1+r))^j C_{t+j} \\ = W_t + \sum_{j=0}^{\infty} (1/(1+r))^j [N_{t+j} - T_{t+j}],$$

which equates the present discounted value of private consumption expenditure to initial asset holdings plus the present discounted value of net of tax labor earnings.<sup>2</sup>

The government sector has a flow budget constraint of the form

$$(4) \quad B_{t+1}/(1+r) - B_t + T_t = G_t,$$

where  $B_t$  = government debt of one-period maturity. Provided that the government debt grows at a rate less than the real rate of return,<sup>3</sup> equation (4) may be utilized to pro-

duce

$$(5) \quad \sum_{j=0}^{\infty} (1/(1+r))^j T_{t+j} \\ = B_t + \sum_{j=0}^{\infty} (1/(1+r))^j G_{t+j}$$

which equates the present discounted value of tax receipts to the initial government debt plus the present discounted value of government purchases.

The representative individual is assumed to be "forward looking" in regard to the fiscal affairs of the government. In particular, the agent recognizes the future tax obligations implicit in current debt issuance, which allows an equivalence between tax or debt finance of a given government expenditure stream. In addition, the individual takes into consideration the benefits to be derived from the future provision of goods and services by the government. Accordingly, the private and public sectors can be integrated by the substitution of the government budget constraint (5) into the representative agent's budget constraint (3) to obtain the following budget constraint in terms of effective consumption:

$$(6) \quad \sum_{j=0}^{\infty} (1/(1+r))^j C_{t+j}^* = (W_t - B_t) \\ + \sum_{j=0}^{\infty} (1/(1+r))^j [N_{t+j} + (\theta - 1)G_{t+j}].$$

Thus, the present discounted value of effective consumption is constrained by the level of net economywide wealth ( $W_t - B_t$ ), plus the present discounted value of labor earnings, plus  $(\theta - 1)$  times the present discounted value of government expenditure. The last term arises because a higher level of government spending imposes a negative (positive) wealth effect on the representative individual as long as  $\theta < (>) 1$ .

The maximization of the individual's objective function (1) subject to the effective intertemporal budget constraint (6) yields as

<sup>1</sup>See, for example, Barro (1981). I ignore other possible channels of influence of government spending on the economy such as providing infrastructure capital as an input to private production processes. On related matters, see my earlier paper (1983) and Willem Buiter (1977).

<sup>2</sup>The solvency condition  $\lim_{k \rightarrow \infty} (1/(1+r))^k w_{t+k} = 0$  has been imposed to obtain equation (3).

<sup>3</sup>To be exact, I impose  $\lim_{k \rightarrow \infty} (1/(1+r))^k B_{t+k} = 0$ . The conditions under which this is likely to hold are discussed by Barro (1974, 1976) and Feldstein (1976).

first-order necessary conditions

$$u'(C_{t+j}^*) = \lambda \cdot ((1+\delta)/(1+r))^j$$

$$j = 0, 1, 2, \dots,$$

along with the intertemporal budget constraint (6). Here,  $\lambda$  is a Lagrangian multiplier attached to (6) in the consumer's maximization problem. The consideration of the choice of consumption in the adjacent periods  $t$ ,  $t+1$  then leads to the Euler equation

$$(7) \quad u'(C_{t+j}^*) = [(1+\delta)/(1+r)]^j u'(C_t^*).$$

Hence, in order for the individual to be choosing an optimal (interior) time path for effective consumption, it must be the case that he cannot improve his welfare standing by reducing effective consumption in one period, say  $t$ , and increasing effective consumption during another period, say  $t+1$ . The cost of reducing effective consumption during period  $t$  and purchasing a bond would be the reduction in utility such an action would entail, or  $u'(C_t^*)/(1+r)$ . The benefit of this action would be the (subjectively discounted) gain in utility during period  $t+1$  to be obtained from the proceeds of the investment, which would be  $u'(C_{t+1}^*)/(1+\delta)$ .

Note the generality of condition (7). For instance, this condition should hold even if utility were also dependent upon leisure (in a manner separable from effective consumption) and there were quantity constraints in the labor market. As long as free access to the credit market is allowed, the agent would allocate resources so as to attain a smooth consumption profile even if during certain periods of his life he faced a situation of involuntary unemployment.

In order to obtain a closed-form solution for consumption, the form of preferences is restricted in the objective function (1). Assuming the momentary utility function is quadratic so

$$u(C_t^*) = -(\bar{C}^* - C_t^*)^2/2,$$

where  $\bar{C}^*$  is the bliss level of effective con-

sumption, the Euler equation is given by

$$(8) \quad C_{t+1}^* = \alpha + \beta C_t^*,$$

where  $\alpha \equiv [(r-\delta)/(1+r)]\bar{C}^*$  and  $\beta \equiv (1+\delta)/(1+r)$ . Using (8) to substitute out  $C_{t+j}^*$  ( $j=1, 2, \dots$ ) and equation (6) allows us to write

$$(9) \quad C_t^* = [(\delta-r)/r(1+r)^2]\bar{C}^* + \left[ \frac{r^2+2r-\delta}{(1+r)^2} \right] \left\{ \sum_{j=0}^{\infty} \left( \frac{1}{1+r} \right)^j [N_{t+j} + (\theta-1)G_{t+j}] + (W_t - B_t) \right\}.$$

Finally, the separation of the period  $t$  levels of income and government spending from the present value term in (9) yields the specification

$$(10) \quad C_t = \beta_0 + \beta_1 N_t + \beta_2 W_t + \beta_3 G_t + \beta_4 T_t + \beta_5 B_t + \beta_6 \sum_{j=1}^{\infty} \left( \frac{1}{1+r} \right)^j N_{t+j} + \beta_7 \sum_{j=1}^{\infty} \left( \frac{1}{1+r} \right)^j G_{t+j},$$

where  $\beta_0 = (\delta-r)\bar{C}^*/[r(1+r)^2]$ ,  $\beta_1 = \beta_2 = -\beta_5 = \beta_6 = \beta_7/(\theta-1) \approx r/(1+r)$ ,  $\beta_3 \approx -(r+\theta)/1+r$ , where the approximations are for  $\delta \approx r$ , and  $\beta_4 = 0$ .

Let us consider the relationship between the specification in equation (10) and that to be found in Feldstein's previous study (1982) of fiscal policy effectiveness:<sup>4</sup>

$$(10') \quad C_t = b_0 + b_1 Y_t + b_2 W_t + b_3 G_t + b_4 T_t + b_5 B_t,$$

<sup>4</sup>I abstract from Social Security wealth although it could be readily incorporated into the analysis. Also, it should be noted that similar criticisms could be made of other studies using the same methodology. Feldstein's was chosen at random.

where all variables are as measured before and  $Y_t$  is permanent income, measured by total national income. Feldstein argues, first, that the Ricardian equivalence theorem implies the restrictions  $b_4 = 0$  and  $b_2 + b_5 = 0$ , which are in accord with the theoretical coefficients in equation (10). Second, he asserts correctly that a test of full fiscal neutrality, whereby current changes in government spending induce an equal, opposite shift in private consumption, would entail the additional restriction that  $b_3 = -1$  in equation (10'). Despite the fact that the proposed restrictions are correct for the test of the null hypothesis, at least three criticisms can be pointed toward this formulation.

The first obvious criticism arises from the treatment of current income as exogenous for the purpose of estimation. As first pointed out by Trygve Haavelmo (1943), since the disturbance term will not be orthogonal to current income, biased and inconsistent estimates will arise in ordinary least squares regression. Feldstein attempts to correct for this bias by employing instrumental variables for income and taxes in some of the reported regressions. Still, the chosen instruments—lagged income and taxes—may not be able to fully eliminate the bias due to serial correlation in the data series. (See Feldstein, 1982, p. 13.)

Second, in his empirical work, Feldstein enters a lagged value of total national income to capture any extra information it may contain as to permanent income. Implicit is the assertion that national income follows a second-order autoregressive process. In this context, it would seem more appropriate to follow Thomas Sargent (1978) by postulating an auxiliary equation for income and computing estimates subject to the cross-equation restrictions between the stochastic processes governing consumption and income which would be imposed by the assumption of rational expectations. Further, the permanent income measure chosen by Feldstein—national income—is inappropriate since it includes future nonlabor income as well as future labor income and the former is already accounted for in the wealth

variable.<sup>5</sup> The appropriate income variable to be chosen is labor income as should be clear from inspection of the specification in (10).

Third, and most important to the issues of the present paper, the formulation in (10') omits any influence of future levels of government spending on current consumption decisions. Feldstein's own fiscal expectations view holds that a change in the current value of government spending or taxes signals future changes in one or both of these variables. This implies that the omitted government spending variables should be expected to be highly correlated with the fiscal policy variables which are included in (10'). Consequently, the coefficient estimates of the latter variables will be biased and will result in incorrect inferences regarding the ability of fiscal policy actions to alter aggregate demand. For example, suppose that current tax revenues are positively correlated with future government spending. Then, since the theory predicts consumption to be negatively related to government spending (assuming

<sup>5</sup>As an illustration of this point, suppose wealth is kept fixed at the value  $W_t$  for  $j = 0, 1, \dots$ , so

$$W_t = \sum_{j=0}^{\infty} (1/(1+r))^j R W_t,$$

where  $R W_t \equiv r W_t / (1+r)$  is the future nonlabor income the individual would receive in each period. Abstracting from governmental variables to center on the particular issue under consideration, we could write permanent income as

$$R \left[ W_t + \sum_{j=0}^{\infty} (1/(1+r))^j N_{t+j} \right]$$

or, by substituting for  $W_t$  in this expression,

$$R \left[ \sum_{j=0}^{\infty} (1/(1+r))^j Y_{t+j} \right],$$

where  $Y_{t+j} \equiv [N_{t+j} + R W_t]$  may be taken to be national income. Thus, Feldstein is being redundant in defining permanent income in terms of national income and adding an independent wealth variable in (9').

$0 < \theta < 1$ ), the omission of future government spending from the regression equation (10') will tend to bias the estimated coefficient on the current tax variable below zero and provide an apparent refutation of the proposition of Ricardian equivalence. In this case, the current tax variable merely acts as a proxy for higher expected government spending. The general point is that, since the current values of the fiscal policy variables carry information regarding future government spending, it is difficult to determine the extent to which the statistical significance of the fiscal policy variables in (10') uncovers a true structural relationship.

The approach of this paper is to abandon the methodology of earlier studies in this area of research and instead to utilize the restrictions which the Euler equation (7) places on the data as in Robert Hall (1978), Marjorie Flavin (1981), N. G. Mankiw, Julio Rotemberg, and Lawrence Summers (1982), Lars Peter Hansen and Kenneth Singleton (1983), and others.<sup>6</sup> This avoids the problems cited above and yields evidence on the substitutability of government spending for private consumption and on the joint hypothesis of rational expectations and Ricardian equivalence.

## II. The Data and Empirical Results

In the study of intertemporal consumption behavior, it is of vital importance to distinguish between consumption and consumer expenditure. At any point in time, consumption might arise without any act of consumer spending (for example, the enjoyment of programs from a previously acquired television set) as well as consumer expenditure without consumption (for example, the purchase of a lawn mower at a winter sale). Ideally, then, one would like to add to current consumer expenditure a flow of services from previously acquired consumer durables and to

subtract current expenditures on durable goods to obtain an adequate measure of consumption. In this paper, an attempt at the latter adjustment is made by defining consumption to be consumer expenditures on nondurables and services. Notice, however, that this adjustment is crude since many goods included in this category would still have durable characteristics. Further, no attempt at the former adjustment is made due to the arbitrariness and difficulties involved in the imputation of a service flow from the stock of consumer durables. Therefore, as the term is used in the empirical analysis below, consumption is per capita consumer expenditure on nondurables and services measured in constant (1972) dollars. Quarterly data are used throughout the study.

The empirical analysis assumes, again, quadratic utility but now in an explicitly stochastic environment so that the Euler equation may be written as

$$(11) \quad E_t C_{t+1}^* = \alpha + \beta C_t^*,$$

where, as before,  $\alpha \equiv [(r - \delta)/(1 + r)]\bar{C}^*$ ,  $\beta = (1 + \delta)/(1 + r)$  and  $E_t$  is the expectations operator conditional on information available up through period  $t$ .

I begin by neglecting the predicted theoretical effect of government spending on the time path of consumption so that equation (11) reduces to Hall's (1978) condition written here as

$$(11') \quad E_t C_{t+1} = \alpha + \beta C_t.$$

Hall estimates (11') and finds that the data support the implication of the permanent income hypothesis that consumption follows a random walk with drift.<sup>7</sup> As a further check

<sup>6</sup>Another interesting avenue would be to follow the instrumental variable approach as in Fumio Hayashi (1982).

<sup>7</sup>The important implication of Hall's model is that  $E(u_t X_{t-1}) = 0$  where  $X_{t-1}$  is any variable in the information set at time  $t-1$  other than  $C_{t-1}$ . To see this, consider a simple Keynesian model with an investment accelerator as  $C_t = cY_{t-1} + e_t$ ;  $I_t = v(Y_t - Y_{t-1}) + \eta_t$ ;  $Y_t = C_t + I_t$ , where  $e_t$  and  $\eta_t$  are both white noise processes. The reduced-form equation for consumption is  $c_t = \beta C_{t-1} + u'_t$ , where  $\beta = (v - c)/(v - 1)$  so for  $v$  large

on the validity of the permanent income hypothesis, Hall in sequence introduces lagged values of consumption, disposable income, and wealth to see if these variables have any predictive power for current consumption apart from that of one-period-lagged consumption. Although past values of wealth—as measured by stock prices—turn out to be statistically significant in predicting current consumption, Hall comes to the overall positive conclusion that "... there is little reason to doubt the life cycle-permanent income hypothesis" (p. 985).

As the present paper is concerned with the impact fiscal policy actions have on the intertemporal path of consumption, consider the effect of past government deficits on current consumption as in the following regression equation:

$$(12) \quad C_t = \alpha + \beta C_{t-1} + \gamma_1 D_{t-1} + \gamma_2 D_{t-2} \\ + \gamma_3 D_{t-3} + \gamma_4 D_{t-4} + u_t,$$

where  $D_t$  is the per capita net deficit of the total government sector measured in constant (1972) dollars. The results from estimating this equation by ordinary least squares for the sample period 1948:I to 1981:IV are listed in Table 1.

The deficit variable makes a statistically important contribution to the predictive power of the equation, with primary influence arising from the first and second lagged values. The  $F$ -statistic for testing the null hypothesis that the coefficients on the deficit variable are all zero is equal to 4.17, substantially above the 5 percent critical value of 2.44 for (4,130) degrees of freedom. Thus, at least at first blush, it appears that a damaging blow has been inflicted upon the Ricardian equivalence hypothesis and hence also upon the theoretical structure of this paper.

However, it may be argued that the influence of government financial variables on

private consumption may be more apparent than real, due to the fact that past taxes or deficits may help to predict current government spending. In this case, if government spending substitutes for private consumption in utility, then the estimates of  $\gamma_1, \dots, \gamma_4$  would be expected to be significantly different than zero. Rather than providing a refutation of the joint hypothesis, the above results could be logically interpreted as evidence in favor of the joint hypothesis, provided that the cross-equation restrictions imposed by the theory cannot be rejected at conventional levels of significance.

So as to take consideration of this point, decompose effective consumption into its private and public components and write from equation (11) the following equation:

$$(13) \quad C_t = \alpha + \beta C_{t-1} + \theta \theta G_{t-1} - \theta G_t^e + u_t.$$

Here,  $G_t^e$  is the expected level of government purchases for time  $t$  conditional upon all information available to the agent at time  $t-1$ . Government spending is measured empirically by per capita government expenditure on goods and services in constant (1972) dollars.<sup>8</sup>

The auxiliary equation to be employed in the prediction of the current level of government spending is given by

$$(14) \quad G_t = \gamma + \varepsilon(L)G_{t-1} + \omega(L)D_{t-1} + v_t,$$

where  $\varepsilon(L) = \sum_1^n \varepsilon_i L^{i-1}$  and  $\omega(L) = \sum_1^n \omega_j L^{j-1}$ ,  $L$  being the lag operator  $LX_t \equiv X_{t-1}$ , and  $v_t$  satisfies the orthogonality condition  $E(v_t | I_{t-1}) = 0$  ( $I_s$  being the information set available to the agent at time  $s$ ) so that  $v_t$  is serially uncorrelated. Written in this form, it is postulated that, apart from past values of government spending, past values of government financial variables

and  $C \approx 1$ ,  $\beta \approx 1$ . However, in the Keynesian case we find that  $E_t(u'_t X_{t-1}) \neq 0$ . Crucial to this argument, of course, is the assumption of time separability in the specification of the consumer's preferences.

<sup>8</sup>The empirical analysis does not attempt to differentiate between government purchases which provide current utility and government purchases which provide future utility either directly or indirectly through private production processes. This distinction was made in the later sections of Kormendi.

TABLE 1—ORDINARY LEAST SQUARES ESTIMATE OF EQUATION (12)  
1948:I TO 1981:IV

Constant	$C_{t-1}$	$D_{t-1}$	$D_{t-2}$	$D_{t-3}$	$D_{t-4}$
1.522 (.754)	.99 (.003)	-.054 (.025)	.066 (.036)	-.042 (.036)	-.026 (.025)
$SER = 2.09$ ; $\bar{R}^2 = .998$ ; $h = .875$ ; $F = 4.17$					

Source: Citibank economic database.

Notes: Estimated standard errors are shown in parentheses.  $h$  is Durbin's test statistic for serial correlation in the residuals in the presence of lagged dependent variables.  $F$  is the value of the statistic appropriate for testing the null hypothesis that the coefficients on the lagged values of the government deficit are all zero.  $C_t$  = per capita consumer expenditure on nondurables and services in constant (1972) dollars.  $D_t$  = per capita net deficit of federal, state and local governments in constant (1972) dollars.

summarized by past deficits help to predict current government expenditure. The linear least squares predictor of  $G_t$  is then given by

$$E_{t-1}G_t \equiv G_t^e = \gamma + \varepsilon(L)G_{t-1} + \omega(L)D_{t-1},$$

which, upon substitution into equation (13), yields the two-equation system below:

$$(15a) \quad C_t = \delta + \beta C_{t-1} + \eta(L)G_{t-1} + \mu(L)D_{t-1} + u_t,$$

$$(15b) \quad G_t = \gamma + \varepsilon(L)G_{t-1} + \omega(L)D_{t-1} + v_t.$$

The "hallmark" of the rational expectations modeling approach is the existence of a set of cross-equation restrictions which is implied by the underlying theoretical structure. In the present case, I obtain

$$(16) \quad \begin{aligned} \delta &= \alpha + \theta\gamma \\ \eta_i &= \begin{cases} \theta(\beta - \varepsilon_i) & i = 1 \\ -\theta\varepsilon_i & i = 2, \dots, n \end{cases} \\ \mu_j &= -\theta\omega_j & j = 1, 2, \dots, m. \end{aligned}$$

Thus, the equation set (16) restricts the way in which past government expenditure and past government deficits may influence present consumption expenditure. In particular, if the Ricardian equivalence proposition does not hold, past values of the government def-

icit should have explanatory power for consumption expenditure apart from their role in forecasting government spending. Consequently, a finding that the data do not do violence to the restriction set (16) yields some ground on which to argue that, to a first approximation, the joint assumption of rational expectations and Ricardian equivalence provides a plausible description of reality.

The empirical procedure is to estimate the system (15), subject to the restrictions (16) by the method of full-information maximum likelihood (FIML) to acquire estimates of the free parameters of the system ( $\alpha, \beta, \theta, \gamma, \varepsilon_1, \dots, \varepsilon_n, \omega_1, \dots, \omega_m$ ), which are  $n + m + 4$  in number. The method allows for nonlinear parameter restrictions within and across equations. The actual estimation was carried out in the TROLL computer package which utilizes the iterative hill-climbing technique developed by Davidon-Fletcher-Powell to maximize the likelihood function. The results of this estimation for the sample period extending from 1984:I to 1981:IV for the case of two lagged values of government spending and two lagged values of the government deficit are reported in Table 2.

Overall, the results appear to be encouraging for the joint hypothesis. Consider first the constrained estimates of the free parameters of the system. The estimated coefficient on the lagged value of consumption is highly significant and equal to unity, with the implication that—holding fixed the level of

TABLE 2—FIML ESTIMATION OF EQUATIONS (15)  
1948:I TO 1981:IV<sup>a</sup>

Constrained	Unconstrained	Hypothesized <sup>b</sup>
$\alpha = 1.360$ (.117)	$\delta = 1.922$ (1.238)	$\delta = .920$
$\beta = 1.002$ (.001)	$\beta = .990$ (.015)	$\beta = 1.002$
$\theta = .229$ (.111)	$\eta_1 = -.024$ (.061)	$\eta_1 = -.088$
$\gamma = 1.293$ (.659)	$\eta_2 = .035$ (.060)	$\eta_2 = .088$
$\epsilon_1 = 1.385$ (.077)	$\mu_1 = -.028$ (.026)	$\mu_1 = -.010$
$\epsilon_2 = -.384$ (.077)	$\mu_2 = -.002$ (.025)	$\mu_2 = -.010$
$\omega_1 = .041$ (.030)	$\tau = 1.267$ (.749)	$\gamma = 1.293$
$\omega_2 = .025$ (.030)	$\epsilon_1 = 1.421$ (.080)	$\epsilon_1 = 1.385$
$\bar{R}_C^2 = .998$	$\epsilon_2 = -.420$ (.080)	$\epsilon_2 = -.384$
$\bar{R}_G^2 = .998$	$\omega_1 = .027$ (.033)	$\omega_1 = .041$
$h_C = 1.17$	$\omega_2 = .026$ (.034)	$\omega_2 = .025$
$h_G = .44$	$\bar{R}_C^2 = .999$	
	$\bar{R}_G^2 = .998$	
	$-2\log_e(L_r/L_u) = 4.281$	

Source: Citibank economic database.

Notes: Estimated standard errors in parentheses. See Table 1 for definition of  $h$ .

<sup>a</sup> $n = m = 2$ .

<sup>b</sup>The coefficients are obtained by substitution of the constrained coefficient estimates into the set of restrictions (16).

government spending—private consumption expenditure follows a random walk process. The point estimate for the substitutability of public spending for private consumption equals .23 and is significantly different from zero at the 5 percent level. This result that government spending substitutes poorly for private spending implies that increases in government purchases of goods and services will have important expansionary effects on aggregate demand even in a setting where the government financing decision is irrelevant to the determination of real variables. In an expanded neoclassical model as in Hall (1978) or Barro (1981), to the extent that such increases in government spending are temporary in nature (for example, wartime expenditure), there would also be a stimulative

effect on real output as the induced rise in the real rate of return would call forth an intertemporal substitution of work effort from the future to the present. Note also that the point estimate of  $\theta = .23$  is roughly in accord with Kormendi's results.

The results indicate that government spending reacts to a process innovation in a cumulative—and borderline unstable—manner. Further, the level of government purchases appears to be positively related to past government deficits, with principal predictive power being confined to the first lagged value of the deficit. Given that government spending and deficits are characterized by positive serial correlation, this latter result may be rationalized along the lines in Barro (1981) or Finn Kydland and Edward Prescott (1980) by the recognition that optimal public finance would require that temporary increases in government spending should be financed by debt creation in an attempt to smooth tax rates over time and, thereby, minimize the deadweight loss due to distortionary labor income taxation.

Next, compare the unconstrained parameter estimates with the hypothesized values obtained by substituting the constrained estimates into the set of restrictions (16). It may be argued in a heuristic manner that the data do not contain substantial evidence against the joint Ricardian equivalence-rational expectations hypothesis if the unconstrained parameter estimates and the hypothesized parameter values do not differ by a substantial amount from one another. Inspection of these two columns in Table 2 indicates that all parameters carry the same signs and are roughly of the same order of magnitude. We should expect, therefore, that a formal statistical test will not lead to a strong rejection of the (joint) null hypothesis.

Turning to this point, since the unconstrained version of the system (15) has  $2(n+m)+3$  regressors and the number of free parameters in the system is equal to  $n+m+4$ , the log-likelihood ratio statistic  $-2\log_e(L_r/L_u)$  is distributed in large samples as a  $\chi^2(k)$  random variable with  $k = [2(n+m)+3] - (n+m+4) = n+m-1$  degrees of freedom, where  $L_r$  is the value of the



TABLE 3—FIML ESTIMATION OF EQUATIONS (15)  
1948:I TO 1981:IV<sup>a</sup>

Constrained	Unconstrained	Hypothesized <sup>b</sup>
$\alpha = 1.370$ (.121)	$\delta = 1.930$ (1.239)	$\delta = 1.068$
$\beta = 1.002$ (.001)	$\beta = .990$ (.002)	$\beta = 1.002$
$\theta = .231$ (.113)	$\eta_1 = -0.26$ (.057)	$\eta_1 = -.093$
$\gamma = 1.308$ (.654)	$\eta_2 = .037$ (.056)	$\eta_2 = .090$
$\epsilon_1 = 1.404$ (.075)	$\mu_1 = -.029$ (.015)	$\mu_1 = -.014$
$\epsilon_2 = -.403$ (.075)	$\gamma = 1.278$ (.150)	$\gamma = 1.308$
$\omega_1 = .061$ (.016)	$\epsilon_1 = 1.442$ (.075)	$\epsilon_1 = 1.404$
$\bar{R}_C^2 = .998$	$\epsilon_2 = -.441$ (.075)	$\epsilon_2 = -.403$
$\bar{R}_G^2 = .998$	$\omega_1 = .049$ (.018)	$\omega_1 = .061$
$h_C = 1.340$	$\bar{R}_C^2 = .999$	
$h_G = .010$	$\bar{R}_G^2 = .998$	
	$+ 2\log_e(L_r/L_u) = 4.280$	

Notes: See Table 2.

<sup>a</sup> $n = 2, m = 1$ .<sup>b</sup>See Table 2.

log-likelihood function under the constrained maximization and  $L_u$  is its value under the unconstrained maximization. A large discrepancy between the constrained and unconstrained values of the log-likelihood function results in a large value of the test statistic and evidence against the null hypothesis that the constrained model is true. For the case examined above where  $n = m = 2$ , the value of the log-likelihood ratio statistic is 4.281, substantially below the 10 percent critical value of the  $\chi^2(3)$  distribution, 6.25 (the implied marginal confidence level is 76 percent, so that the null hypothesis cannot be rejected at a significance level lower than 24 percent). Therefore, in this case, the data are incapable of rejecting the null hypothesis at conventional significance levels.

Notice that the least significant of the variables in the constrained estimation is the second lagged value of the deficit in the government purchases equation. A natural course would be to reestimate the model for the case of two lagged values of government spending and one lagged value of the govern-

ment deficit. The results of this estimation over the sample period 1948:I to 1981:IV are listed in Table 3. The constrained coefficient estimates maintain the same (or nearly the same) values and levels of statistical significance. The only exception is the first lagged deficit variable, the coefficient of which experiences an increase in its value and a fall in its estimated standard error. Again, the unconstrained parameter estimates and the values implied by the constrained estimates and the restriction set (16) always have the same signs and are similar in magnitude. In this case, the log-likelihood ratio statistic is distributed asymptotically as  $\chi^2(2)$  and takes on the value 4.280 which is still below the 10 percent critical value of the  $\chi^2(2)$  distribution, 4.61 (the implied marginal confidence level is 87 percent). Although the elimination of the two-period-past deficit raises the confidence level at which the null hypothesis can be rejected, it remains impossible to argue that the data provides evidence against the joint proposition of Ricardian equivalence and rational expectations at conventional levels of significance.

The values of the Durbin  $h$ -statistic in Tables 2 and 3 do not indicate the presence of a significant amount of autocorrelation in the estimated residuals of the consumption and government purchases equations. Nevertheless, a further check on the robustness of the results obtained above was implemented by expanding the lags of the government spending and deficit variables. Table 4 reports values of the associated log-likelihood ratio statistics as well as estimated values of the substitutability parameter,  $\theta$ .<sup>9</sup> In all cases, the null hypothesis under consideration cannot be rejected at the 10 percent level of significance. In one case,  $n = m = 6$ , it would be impossible to reject the null hypothesis at a significance level lower than 25 percent. The value of the substitutability parameter tends to rise with the number of included lags of government spending and deficits, to as much as .421 with the inclusion of two years of past spending and deficits. For the

<sup>9</sup>Complete results are available from the author by request.

TABLE 4—VALUES OF THE LOG-LIKELIHOOD RATIO STATISTICS AND  $\theta$  ESTIMATES

$n = m =$	$N$	$\theta$	$-2\log_e L_r/L_u$	$k$	$\chi^2$		
					.75	.90	.95
3	136	.331	8.710	5	6.63	9.24	11.07
4	136	.332	11.481	7	9.04	12.02	14.07
6	134	.332	13.561	11	13.70	17.28	19.68
8	132	.421	21.568	15	18.20	22.31	25.00

Notes:  $N$  = sample size,  $k$  = degrees of freedom for likelihood ratio test.

cases  $n = m = 3$ ,  $n = m = 4$ , and  $n = m = 6$ , however, the substitutability parameter takes on the value  $\theta = .33$ , almost exactly the value found by Kormendi utilizing the conventional methodology. On the basis of these results we may come to the overall conclusions that the data do not appear capable of strongly rejecting the Ricardian equivalence theorem and that rises in government spending will only induce a partial *ex ante* crowding out of private consumption.

### III. Conclusion

This paper has investigated the question of fiscal impotence within an explicit rational expectations optimizing framework. Two questions were posed: To what extent does government spending induce an *ex ante* crowding out of private consumption expenditure? To what degree do the data contain evidence against the tax discounting hypothesis associated with the theoretical analysis of Barro (1974)? Public expenditure was seen to reduce private consumer expenditure on nondurables and services in the range of 23 to 42 percent, a range which is compatible with Kormendi's results. The values of the log-likelihood ratio statistics are too low to reject the joint rational expectations-Ricardian equivalence hypothesis at the typical 5 or 10 percent levels, a finding which adds some support to the earlier empirical work of Barro (1978), Kochin, Charles Plosser (1982), and Kormendi.

The new-classical school of macroeconomic policy stresses the real effects of government spending rather than the method by which such spending is financed. The primary effect of temporary increases in

government spending on output arise from the attempt by economic agents to smooth effective consumption levels over time. Hence, if  $0 < \theta < 1$ , this attempt will induce a reallocation of resources from other periods to the present which, in turn, will increase rates of return and cause an intertemporal substitution of work effort and a contemporaneous expansion of output. The empirical results of this paper suggest that this view of the effects of fiscal policy actions on the economy deserves at least some credibility, perhaps even the status of the working hypothesis.

### APPENDIX

The variable definitions and statistical sources are:

$C_t$  = real per capita consumer expenditure on nondurable goods and services;

$D_t$  = net real per capita deficit of federal, state and local governments;

$G_t$  = real per capita expenditures of federal, state and local governments.

Deflation of nominal aggregates is by the implicit price deflator (1972 = 100) and total population of the United States. All variables are taken from the Citibank economic database, "Citibase."

### REFERENCES

- Aschauer, David A., "A Theory of Crowding Out," unpublished, 1983.
- Bailey, Martin J., *National Income and the Price Level*, New York: McGraw-Hill, 1971.
- Barro, Robert J., "Are Government Bonds Net Wealth?," *Journal of Political Economy*,

- November/December 1974, 82, 1095-117.
- \_\_\_\_\_, "Reply to Feldstein and Buchanan," *Journal of Political Economy*, April 1976, 84, 343-50.
- \_\_\_\_\_, *The Impact of Social Security on Private Saving*, Washington: American Enterprise Institute, 1978.
- \_\_\_\_\_, "Output Effects of Government Purchases," *Journal of Political Economy*, December 1981, 89, 1086-121.
- Buiter, Willem, "Crowding Out and the Effectiveness of Fiscal Policy," *Journal of Public Economics*, June 1977, 7, 309-28.
- Feldstein, Martin, "Perceived Wealth in Bonds and Social Security," *Journal of Political Economy*, April 1976, 84, 331-36.
- \_\_\_\_\_, "Government Deficits and Aggregate Demand," *Journal of Monetary Economics*, January 1982, 9, 1-20.
- Flavin, Marjorie A., "The Adjustment of Consumption to Changing Expectations about Future Income," *Journal of Political Economy*, October 1981, 89, 974-1005.
- Haavelmo, Trygve, "The Statistical Implications of a System of Simultaneous Equations," *Econometrica*, January 1943, 11, 1-12.
- Hall, Robert, "Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence," *Journal of Political Economy*, December 1978, 86, 971-87.
- \_\_\_\_\_, "Labor Supply and Aggregate Fluctuations," *Carnegie-Rochester Conference Series on Public Policy: On the State of Macro-Economics*, Spring 1980, 12, 7-33.
- Hansen, Lars Peter and Singleton, Kenneth J., "Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns," *Journal of Political Economy*, April 1983, 91, 249-64.
- Hayashi, Fumio, "The Permanent Income Hypothesis: Estimation and Testing by Instrumental Variables," *Journal of Political Economy*, October 1982, 90, 895-916.
- Kochin, Lewis, "Are Future Taxes Discounted by Consumers? Comment," *Journal of Money, Credit and Banking*, August 1974, 6, 385-94.
- Kormendi, Roger, "Government Debt, Government Spending and Private Sector Behavior," *American Economic Review*, December 1983, 73, 994-1010.
- Kydland, Finn and Prescott, Edward, "A Competitive Theory of Fluctuations and the Feasibility and Desirability of Stabilization Policy," in S. Fischer, ed., *Rational Expectations and Economic Policy*, Chicago: University of Chicago Press, 1980.
- McCallum, Bennett, "Topics Concerning the Formulation, Estimation and Use of Macroeconometric Models with Rational Expectations," *Proceedings*, American Statistical Association: Business and Economics Statistics Section, 1979.
- Mankiw, N. G., Rotemberg, Julio and Summers, Lawrence, "Intertemporal Substitution in Macroeconomics," Working Paper, Massachusetts Institute of Technology, 1982.
- Nelson, Charles, "Rational Expectations and the Estimation of Econometric Models," *International Economic Review*, October 1975, 16, 555-61.
- Plosser, Charles, "Government Financing Decisions and Asset Returns," *Journal of Monetary Economics*, May 1982, 9, 325-52.
- Sargent, Thomas, "Rational Expectations, Econometric Exogeneity, and Consumption," *Journal of Political Economy*, August 1978, 86, 673-700.
- Tanner, J. Ernest, "An Empirical Test of the Extent of Tax Discounting," *Journal of Money, Credit and Banking*, May 1979, 11, 214-18.

# General Equilibrium Computations of the Marginal Welfare Costs of Taxes in the United States

By CHARLES L. BALLARD, JOHN B. SHOVEN, AND JOHN WHALLEY\*

In recent years, increasing attention has been paid by public finance economists to the marginal excess burden (*MEB*)<sup>1</sup> per additional dollar of tax revenue. Estimates of *MEBs* stand in contrast to estimates of the welfare cost of taxes which are calculated by totally removing existing taxes and replacing them with equal yield lump sum taxes. Instead, an *MEB* estimate measures the incremental welfare costs of raising extra revenues from an already existing distorting tax. Earlier estimates of *MEBs* have either concentrated on particular portions of the tax system, or have employed partial equilibrium methods. Here, we examine the *MEB* of all major taxes in the United States, using a multisector, dynamic computational general equilibrium model. This allows us to calculate simultaneously the marginal welfare effects of individual income taxes, corporate taxes, payroll taxes, sales and excise taxes, and other smaller sources of revenue.

We find that the marginal excess burden of taxes in the United States is large. The welfare loss from a 1 percent increase in all distortionary tax rates is in the range of 17 to 56 cents per dollar of extra revenue, when we use elasticity assumptions that we consider to be plausible. Consequently, a public proj-

ect must produce marginal benefits of more than \$1.17 per dollar of cost if it is to be welfare improving. This suggests that many projects accepted by government agencies in recent years on the basis of cost-benefit ratios exceeding unity might have been rejected if the additional effects of distortionary taxes had been taken into account. The cost-benefit standard should be more stringent. Another implication of our results is that a tax reform that lowers tax rates by a relatively small amount might significantly reduce the total welfare costs of taxes.

We also calculate the marginal excess burden from increases in various parts of the tax system. Not surprisingly, we find that the *MEB* for a given part of the tax system is greater when the taxed activity is assumed to be more elastic. The *MEB* from capital taxes responds a great deal to the saving elasticity and the *MEB* from labor taxes responds a great deal to the labor supply elasticity. In general, it appears that the *MEBs* are greater for activities which face high or widely varying tax rates. These conclusions are, in general, in accord with those drawn from a simple, partial equilibrium model (see Edgar Browning). Such a model indicates that *MEBs* would be proportional to the elasticity of the taxed activity and proportional to the tax rate.

It is worthwhile to explain our treatment of public goods and the precise tax replacement experiment implicit in our calculations. The literature on the optimal provision of public goods due to Paul Samuelson (1954), Peter Diamond and James Mirrlees (1971a, b), Partha Dasgupta and Joseph Stiglitz (1972), and A. B. Atkinson and N. H. Stern (1974) sets out the conditions for the optimal quantity of a pure public good. Atkinson and Stern modify Samuelson's conditions to account for the excess burden of distortionary taxes used to finance public

\*Departments of Economics: Michigan State University, East Lansing, MI 48824; Stanford University, Stanford, CA 94305; University of Western Ontario, London, Ontario, N6A 5C2 Canada, respectively. An earlier version of this paper appeared as NBER Working Paper No. 1043. We thank Don Fullerton, Larry Martin, Paul Menchik, Joel Slemrod, Charles Stuart, and an anonymous referee for helpful comments, and Janet Stotsky for research assistance. This work was supported by the NBER Project on the Government Budget and the Private Economy and by the U.S. Treasury Department. Any errors are ours.

<sup>1</sup>For example, Edgar Browning (1976), Harry Campbell (1975), Ingemar Hansson and Charles Stuart (1983), and Dan Usher (1982).

goods provision. Although Atkinson and Stern are not concerned with calculating *MEBs* as such, their work is closely related to ours. They allow for complementarity between public goods and private goods, and, if complementarity is sufficiently great, their model can call for an even greater level of public goods than the simple Samuelson model. Our model does not allow for such complementarity since public goods do not enter household utility functions in our framework. In our model, the government uses its revenues to provide transfer payments to the household sector, and it makes exhaustive expenditures that do not directly affect consumer utility or the structure of production. If we were to extend our model to account for complementarity, our measure of *MEB* might be reduced.

Regarding the type of tax change experiment we undertake, it is worthwhile emphasizing that the questions we ask are *not* in the realm of what Richard Musgrave (1959) calls "differential incidence." Studies of differential incidence (including previous studies involving this model, such as Don Fullerton et al., 1981, and Fullerton, Shoven, and Whalley, 1983) hold constant the size of the government. When a distortionary tax is increased, there is an offsetting rebate. In this paper, we analyze what Musgrave would term "balanced budget incidence." We raise distortionary taxes and the government in the model uses the additional revenue for exhaustive expenditures. There is no lump sum rebate to consumers. The foregone alternative is a lower level of taxation rather than a lump sum tax.

#### I. A General Equilibrium Model of the U.S. Economy and Tax System: Structure and Data

To keep the focus of this paper on results and policy implications, only a brief overview of model structure is given here. We provide a detailed description of our model in chapters 3–7 of Ballard et al. (1985). First, we summarize the production side of the model. In any single period, there are 19 producer-good industries that use capital and labor in constant elasticity of substitution (*CES*) value-added functions. They also use

the outputs of other industries through a matrix of fixed input-output coefficients. The tax rates on labor for each industry are derived by taking payroll taxes and other contributions as a proportion of labor income, while the tax rates on capital for each industry are derived by taking corporate income, corporate franchise, and property taxes as a proportion of capital income. Each of these 19 producer goods is used directly for investment, for net exports, and for exhaustive government expenditures. In any period, consumers allocate their consumption among 15 consumer goods. The transformation of producer goods into consumer goods is represented by a matrix of fixed coefficients. This procedure is necessary because the goods classification of consumer expenditure data is different from the classification of the outputs of the 19 production sectors.

On the consumer side of the model, we have 12 consumer groups, which are distinguished by their money income<sup>2</sup> in 1973 (the basic data year for the model). Each consumer group has an initial endowment of capital and labor. Consumer decisions regarding factor supplies are made jointly with their consumption decisions. Each household at any point in time has a nested *CES* utility function of the form:

$$(1) \quad U = U \left[ H \left( \prod_{i=1}^{15} X_i^{\lambda_i}, l \right), C_f \right],$$

where  $H$  is the instantaneous utility function defined over current consumption commodities  $X_i$  and leisure  $l$ , and the function  $U$  determines the allocation between current welfare and expected incremental future consumption,  $C_f$ . The 15 current consumption commodities  $X_i$  are aggregated using a Cobb-Douglas function, whereas both  $U$  and

<sup>2</sup> These are incomes as defined for the 1973 *Consumer Expenditure Survey*. Money incomes exclude imputed income from home ownership, and sheltered capital income of various kinds. Even though we differentiate among consumers with this restricted definition of income, we do impute all kinds of capital income to every consumer group. Thus, each consumer group's income in our model calculations is greater than its narrowly defined money income.

$H$  are CES functions. Consumers are infinitely lived, so that there are no bequests.

The government collects taxes from the production and demand sides of the economy and uses the revenue in a balanced budget. The government purchases producer goods, makes direct transfer payments to consumers, and subsidizes government enterprises. A simple formulation of international trade closes the model.

In this model, we calculate a dynamic sequence of static equilibria. In essence, we examine a series of single-period equilibria, sequenced through saving decisions which change the time profile of the economy's capital stock. Saving in each period depends on the expected rate of return on saving in future periods. The simulations reported here use the assumption of myopic expectations.<sup>3</sup> Because of this assumption, the current period rate of return on capital and other current prices are all that we require to solve the utility-maximization problem for each household. With myopic expectations, the price of expected future consumption varies inversely with the current period rate of return, which consumers expect will apply to all future periods. Maximizing  $U$ , subject to a budget constraint, gives the desired level of  $C_t$  for each consumer. The demand of  $C_t$  is then translated into a demand for saving in the current period. The latter is, in turn, translated into a vector of investment demands for the 19 industry outputs.

We specify our model by calibrating to the same benchmark equilibrium data set for 1973 that is used in Ballard et al. A full updating of the data set to a more recent year would be costly and has not been done for our calculations; most of the main features of the U.S. tax system have not changed greatly in the last decade. However, the marginal rate of taxation of corporate capital income may now be lower than our data and techniques suggest. The data set uses five major sources. These are the 1973 Depart-

ment of Labor *Consumer Expenditure Survey*, the July 1976 *Survey of Current Business*, the Bureau of Economic Analysis Input-Output Matrix, unpublished worksheets of the U.S. Department of Commerce National Income Division, and the U.S. Treasury Department's Merged Tax File. In order to generate a consistent data set, a number of adjustments are made. All data on industry and government uses of factors are accepted as given, while the data on consumer factor incomes and expenditures are correspondingly adjusted. Tax receipts, transfers, and government endowments are accepted as given, and government expenditures are adjusted in order to yield a balanced budget. Similar adjustments ensure that supply equals demand for all goods and factors, and that trade is balanced.

The fully consistent data set defines a single-period benchmark equilibrium in transactions terms. These observations on values are then separated into prices and quantities by assuming that a physical unit of a good or factor is the amount that sells for one dollar. All benchmark equilibrium prices are thus \$1, and the observed values are the benchmark quantities.

The equilibrium conditions of the model are then used to determine the behavioral equation parameters consistent with the benchmark data set. This procedure calibrates the model to the benchmark data, in the sense that the benchmark data can be reproduced as an equilibrium solution to the model before any policy changes are considered. In order to implement this procedure, we specify the elasticities of substitution between capital and labor in each industry on the basis of econometric estimates in the literature. We also specify labor supply and saving elasticities (also based on literature sources), to which substitution elasticities in preferences are calibrated. Factor employments by industry are used to derive production function weights, and expenditure data are used to derive utility function weights. This calibration procedure ensures that, given the benchmark data, the various agents' behaviors are mutually consistent before we evaluate policy changes.

The elasticities of labor supply and saving are especially important parameters for our

<sup>3</sup>We have investigated the sensitivity of our results with respect to changes in the assumptions about expectations, using the procedure developed by Ballard and Lawrence Goulder (1982). We find that different expectational structures have little effect on the results.

results. There are a large number of estimates for the uncompensated elasticity of labor supply with respect to the real, net-of-tax wage. Elasticity estimates for males are mostly small and negative, ranging from  $-0.40$  to zero. George Borjas and James Heckman (1978) review these econometric studies and suggest a range between  $-0.19$  and  $-0.07$ . The estimates for females are more often positive, and can be large in absolute value. Mark Killingsworth (1982) finds that the elasticity estimates for females are mostly between  $0.20$  and  $0.90$  in cross-section studies. We use three values for the uncompensated labor supply elasticity. A value of  $0.15$  is our central estimate, and we also use elasticities of  $0.0$  and  $0.30$ . We calibrate these values by specifying the elasticity of substitution between present consumption and present leisure for the  $H$  function in equation (1) for each consumer.

The other key parameter is the elasticity of saving with respect to the real, after-tax rate of return, which we use to determine values for the elasticity of substitution between present consumption,  $H$ , and future consumption,  $C_f$ , for each consumer in the model.

There is considerable literature controversy regarding the value of the uncompensated saving elasticity. For a long time, the consensus appeared to favor a zero value for this elasticity, a proposition termed Denison's Law, after Edward Denison (1958). In more recent work, Michael Boskin (1978) has estimated this elasticity to be approximately  $0.3$  to  $0.4$ , although Lawrence Summers (1981) has shown that reasonable parameter values in life cycle model may imply saving elasticities between  $1.5$  and  $3.0$ . Each of these studies has problems of interpretation. In particular, for reasons outlined in the paper by David Starrett (1982), Summers's elasticity figures may be high. We focus on simulations using the values of  $0.0$  and  $0.4$  for the saving elasticity. We also consider a high value of  $0.8$ . As might be expected, the marginal excess burden estimates increase as the saving elasticity increases. However, the labor supply elasticity seems to be the more important parameter.

The value used for the real net-of-tax return to capital in the benchmark data is

important, since this value is used to calibrate preference parameters under the assumption of intertemporal utility maximization. It also determines the rate of time preference in the benchmark sequence of equilibria. We use  $4$  percent for the average value of this parameter, but each income class receives a net-of-tax return that depends on its own marginal tax rate.

The dynamic behavior of the model depends on the steady-state growth rate assumed for the benchmark equilibrium sequence. To derive this rate, we compare the amount of observed 1973 saving to the capital stock. This gives us a growth rate of the capital service endowment of  $2.89$  percent per year. We assume that labor (in effective units) grows at the same rate. Though labor endowments grow at this fixed annual rate in both the benchmark sequence and the revised sequence, the demand for leisure is endogenous. This implies that market labor supply growth may differ when prices change until balanced growth is reestablished. Though the capital stock grows at this rate in the benchmark sequence, endogenous saving implies that, in the revised case, capital services may grow at a different rate. The  $2.89$  percent labor growth rate is assumed to be equally divided between Harrod-neutral technical change and population growth. Our welfare measures of tax changes are adjusted to account only for the initial population size.

## II. Model Treatment of Taxes

The model incorporates each of the major taxes in the United States. In Table 1, we outline how these are modeled; summary information on the tax rates in the model is presented in Table 2. Mean factor and consumer tax rates across industries and commodities are reported, with indications of the dispersion in tax rates.

The treatment of each tax in the model reflects assumptions we make about the operation of the tax system. Thus, we combine the corporate and property taxes to produce an overall tax rate on capital income originating in each industry. We define these capital tax rates as a proportion of net-of-tax income; thus tax rates can exceed unity. The

TABLE 1—U.S. TAXES AND THEIR TREATMENT IN THE MODEL

Tax	Treatment in the Model	Difficulties of Model Treatment
1. Corporate taxes (including state and local) and corporate franchise taxes	Ad valorem tax on use of capital services by industry	Some argue for treatment as a lump sum tax; model treatment ignores role of financial instruments
2. Property taxes	Ad valorem tax on use of capital services by industry	Differential rates across jurisdictions ignored
3. Social Security taxes, Unemployment Insurance and Workmen's Compensation	Ad valorem tax on use of labor services by industry	Benefit related nature of contributions; arbitrary distinction between public and private insurance programs
4. Motor vehicles tax	Ad valorem tax on use of motor vehicles by producers	In practice, a yearly registration fee and not a purchase tax; averaging over jurisdictions
5. Retail sales taxes	Ad valorem taxes on purchase of consumer goods	Averaging of rates over states
6. Excise taxes	Ad valorem taxes on output of producer goods	Taxes often expressed as charge per unit physical measure such as volume
7. Other indirect business taxes and nontax payments to government	Ad valorem tax on output of producer goods	Payments depend on output levels by industry to only limited extent; averaging of rates over states
8. Personal income taxes (including state and local)	Linear function for each consumer where tax on capital affects industry allocation; 30 percent of savings currently deductible	Detailed deductions and exemptions not specifically considered in model

average tax rate on capital income at the industry level is about 0.97, which corresponds to a tax rate on gross capital income of just under 50 percent.<sup>4</sup>

In modeling the corporate tax, we follow the tradition of Arnold Harberger (1962, 1966) who treats it as a partial factor tax, even though more recently this has been the subject of active debate. Stiglitz (1973), for instance, has argued that if all marginal investments by firms are debt financed, the corporate tax operates as a lump sum tax. However, many features of corporate financial behavior remain unexplained, and we follow Harberger's procedure of treating the corporate tax as an ad valorem tax on capital, with average and marginal tax rates the same.

<sup>4</sup>These rate estimates do not incorporate the reductions in capital tax rates which were part of the 1981 Economic Recovery Tax Act and the further changes of the 1982 Tax Equity and Fiscal Responsibility Act. For a study of the effects of these changes in tax rates, see Fullerton and Yolanda Henderson (1983).

As a result, differences in capital income tax rates cause capital to be misallocated across industries.

In addition, the corporate tax affects saving decisions, since savers who acquire corporate equity indirectly pay these taxes on the return to their savings. Further distortions operate through the tax treatment of depreciation. While depreciation allowances operate at rates that are faster than true depreciation, they are calculated on a historical cost basis. Capital tax rates also include the investment tax credit. All these features combine to produce a pattern of tax rates by industry which is discriminatory.

A further key feature of our specification of capital tax rates is the assumption that average and marginal tax rates are the same. Fullerton (1984) has suggested a number of reasons why marginal and average rates need not be the same, and argues that under current laws, marginal rates are probably lower than average rates. Consequently, our specification may overstate marginal excess bur-



TABLE 2—LEVEL AND DISPERSION OF TAX RATES IN THE MODEL

Type of Tax	Sectors on Which Tax Is Levied	Weights	Weighted Tax Rate Statistics		
			Mean of Marginal Tax Rates	Standard Deviation	Coefficient of Variation <sup>b</sup>
Capital Taxes at Industry Level	19 Industries	Capital Use	0.970	0.729	0.752
Labor Taxes at Industry Level	19 Industries	Labor Use	0.101	0.009	0.092
Consumer Purchase Taxes	15 Goods	Total Consumption	0.067	0.140	2.101
Output Taxes	19 Industries	Output	0.008	0.035	4.612
Motor Vehicle Taxes	Intermediate Use of Motor Vehicles in 19 Industries	Use of Motor Vehicles	0.052	0.051	0.992
Personal Income Taxes <sup>a</sup>	12 Consumer Groups	Income	0.239	0.101	0.424

<sup>a</sup>Personal income tax rates are expressed as a proportion of gross income, whereas the other rates are expressed as proportions of net-of-tax capital income by industry, labor income by industry, etc.

<sup>b</sup>Coefficients of variation will not equal the quotients of the corresponding standard deviations and means because of rounding in the standard deviations and means.

dens as far as this portion of the tax system is concerned. The assumption of equality between marginal and average rates is less contentious in the case of tax rates on labor at the industry level, which we calculate using data on Social Security and other contributions.

We treat the property tax as a differential tax on capital by sector (similarly to the corporate tax). This falls most heavily on residential housing, but structures in other capital-using industries in the economy are also liable for the tax. As with the corporate income tax, both static and dynamic distortions occur.

Income tax rates differ substantially among consumers, with each of the 12 consumer groups facing a linear income tax schedule. Marginal tax rates rise from 0.01 for the poorest group to 0.41 for the richest.

The key distortions caused by the income tax affect factor supply decisions. It is widely recognized that the income tax distorts labor supply. In addition, the supply of new capital through saving is affected (by the "double" taxation of saving), although these effects are partially offset by the tax treatment of pensions and housing. We assume that 30 percent of saving is sheltered in this way. (This assumption is based on calculations using the 1976 *Flow of Funds Accounts*.) However, since saving is heavily concentrated in the

top tail of the income distribution, much of the saving in the economy occurs where the tax rates are highest.<sup>5</sup>

In addition to distorting factor-supply decisions, the income tax also has important features which distort choices among industries and commodities. The most prominent of these is the preferential treatment of housing that results from the absence of tax on the imputed income of owner-occupied housing. This is compounded by the preferential treatment for capital gains on houses.

Consumer sales and excise tax rates average about 6.7 percent in the model, and rates for most goods are reasonably low. There are three notable exceptions: the tax on alcoholic beverages is 87.5 percent, on tobacco, 95.8 percent, and on gasoline and other fuels, 29.5 percent.

Consumer sales taxes have a variety of effects. Even if the sales tax system covers all commodities evenly, it still distorts labor supply decisions. Additional distortions come from the nontaxation of food and other exempted items. Also, the specific excises on alcohol, tobacco, and gasoline are sharply

<sup>5</sup>Our model exaggerates this effect, since we do not capture life cycle differences among households. However, the evidence provided by Paul Menchik and Martin David (1983) indicates that lifetime saving is also concentrated.

discriminatory in our model, since we treat them (along with sales taxes) as ad valorem taxes. We recognize that this latter treatment is contentious. The taxes on alcohol and tobacco could be defended as Pigovian externality-correcting taxes. The gasoline tax is often viewed as a benefit-related fee for the use of the highway system. Because of these considerations, and because our formulation of the consumer's utility function may overstate the elasticity of demand for these products, we report two sets of results for the *MEB* from increases in consumer sales taxes. In the first, we evaluate the effect of an increase in the tax rate on every commodity. In the second, we raise only the tax rates on commodities other than alcohol, tobacco, and gasoline.

### III. Use of the Model in *MEB* Calculations

In our discussion of the various types of taxes, we have distinguished intertemporal distortions (that affect saving decisions) from intersectoral distortions (that affect allocations among industries or consumer goods). Many of the general equilibrium models that exist today can calculate only a single equilibrium. Consequently, they are poorly equipped to analyze the relative importance of intertemporal and intersectoral distortions. Our model allows us to assess intertemporal distortions as well as intersectoral ones. We calculate a sequence of equilibria, covering an arbitrarily long period of time. The equilibria are connected by endogenous saving decisions and exogenous growth of labor endowments.

In each single-period equilibrium, utility-maximizing consumers and profit-maximizing producers reach a competitive equilibrium where all profits are zero and supply equals demand for each good and factor. We use a variant of the Factor Price Revision Rule recently developed by Larry Kimbell and Glenn Harrison (1984) to calculate prices that satisfy these conditions for each time period. Although this algorithm is not guaranteed to converge, we have encountered no convergence problems, and this procedure is substantially faster than O. H. Merrill's (1972) algorithm, which we have used in earlier work on this model.

In each single-period equilibrium, markets are perfectly competitive, and there is no involuntary unemployment of factors, nor are there any externalities, quantity constraints, or barriers to factor mobility. The first equilibrium in the benchmark sequence replicates the 1973 equilibrium data set. In the no-policy-change case, subsequent equilibria are scaled-up versions of the initial equilibrium due to the balanced growth assumption. Prices remain constant, and all quantities grow at the same rate (the exogenous rate of growth of the effective labor force). When we alter tax parameters, we calculate a revised sequence of equilibria by computing a complete set of prices and quantities for each equilibrium in the sequence under an alternative tax policy. We estimate the changes in utility and income for each consumer group, changes in national income, and all new factor allocations among industries between pairs of comparable equilibria in the old (no-policy-change) and revised (after-policy-change) sequences.

Since we cannot compute an infinite sequence of equilibria, we calculate equilibria for a preselected number of years and then use a termination term. The welfare evaluation of the termination term is only correct if the economy is on a steady-state growth path, as is the case in our base-case sequence of equilibria. In a revised-case sequence, the tax change generates a transitional path that approaches a new steady-state growth path and the termination term will only be approximately correct. The accuracy of this approximation becomes better as the economy approaches the new steady-state growth path. In our calculations of marginal excess burden, the changes in relative prices are small since the tax changes are small. We calculate an extremely close approximation by computing our equilibria 100 years into the future. These are spaced five years apart, giving us a sequence of 21 equilibria.

In earlier applications of our model, the government spends any extra tax revenues it receives on both additional transfer payments to consumers and purchases of goods and factors. However, it is easier to interpret the results of this paper if transfer payments do not change. Consequently, for the results reported here, we have changed the model

such that the level of real transfer payments of each consumer group remains the same in each period of the revised-case equilibrium sequence compared to the corresponding period of the base-case sequence.

Our objective is to compare the dollar value of the loss of consumer welfare resulting from an increase in distortionary taxes with the amount of revenue that the tax increase generates. For the loss of consumer welfare, we calculate the present value of a stream of Hicksian equivalent variations. Each of these is calculated using contemporaneous utility in comparable base and revise equilibrium calculations, as described in chapter 7 of Ballard et al. It should be noted that our consumer utility functions do not incorporate public goods. The implicit assumption is that public goods enter utility in a separable manner. We want to compare the dollar value of the loss in consumer utility from leisure and goods due to the tax with the revenue collected by the increase in the tax.

In order to get a similar present value figure for the change in revenue, we correct for changes in relative prices over time, since the dollar increase in revenue in one period is not strictly comparable with the dollar increase in revenue in another period. The model assumption is that government purchases of goods and factors are characterized by constant expenditure shares. Instead of using a Laspeyres price index or some other index to correct for relative price changes, we use the expenditure function associated with the implicit Cobb-Douglas utility function of the government. A different assumption about the pattern of government expenditure could alter the results, since the government does not spend marginal tax revenue in the same way that consumers would have spent it if it had been returned to them in lump sum form.

#### IV. Results

The marginal excess burden calculations produced by the model are shown in Tables 3 and 4. Table 3 shows the *MEB* from raising all marginal tax rates by 1 percent for different saving and labor supply elasticities. Table 4 reports *MEB* estimates from raising

TABLE 3—MARGINAL EXCESS BURDEN PER ADDITIONAL DOLLAR OF REVENUE FOR U.S. TAXES

Labor Supply Elasticity	Saving Elasticity		
	(i) 0.0	(ii) 0.4	(iii) 0.8
(i) 0.0	.170	.206	.238
(ii) 0.15	.274	.332	.383
(iii) 0.30	.391	.477	.559

additional revenue from alternate portions of the tax system for different elasticity configurations.

Estimates of marginal excess burdens in Table 3 are substantial. They indicate that the transfer of an additional dollar to the government causes a deadweight loss in the range of 17 to 56 cents. This means that additional public expenditures ought to be undertaken only if their marginal benefits are at least 17 percent greater than the revenues needed to fund the project, if it has to be financed by additional distorting taxes.

As might be expected, marginal excess burdens are greater when higher elasticity values are used. The results are more sensitive to changes in the uncompensated labor supply elasticity than to changes in the saving elasticity. We would place the most confidence in our estimates using the middle elasticities (.4 and .15). An uncompensated saving elasticity of 0.8 and an uncompensated labor supply elasticity of 0.3 have been added to Table 3 mainly to illustrate the sensitivity of the results to changes in these parameters.

In Table 4, we report *MEB* estimates for cases where additional revenues are raised through the major tax subgroups. For these cases, we only use the labor supply elasticities of 0.0 and 0.15 and saving elasticities of 0.0 and 0.4 as parameter value combinations.

For the most part, the various parts of the tax system do not generate vastly different *MEBs*. However, we can generally say that the more elastic activities have higher *MEBs*. With a saving elasticity of 0.4, capital taxes lead to the greatest *MEBs*. With a labor supply elasticity of zero, the labor taxes at the industry level cause relatively small amounts of marginal distortion. If we focus on our central case (with elasticities of 0.4

TABLE 4—MARGINAL EXCESS BURDEN FROM RAISING EXTRA REVENUE FROM SPECIFIC PORTIONS OF THE TAX SYSTEM

Uncompensated Saving Elasticity:	0.0	0.4	0.0	0.4
Uncompensated Labor Supply Elasticity:	0.0	0.0	0.15	0.15
All Taxes	.170	.206	.274	.332
Capital Taxes at Industry Level	.181	.379	.217	.463
Labor Taxes at Industry Level	.121	.112	.234	.230
Consumer Sales Taxes	.256	.251	.384	.388
Sales Taxes on Commodities other than Alcohol, Tobacco, Gasoline	.035	.026	.119	.115
Income Taxes	.163	.179	.282	.314
Output Taxes	.147	.163	.248	.279

for saving and 0.15 for labor supply), we see that capital taxes, consumer sales taxes, and income taxes cause the greatest distortion, followed by output taxes and labor taxes at the industry level. This is almost exactly the same ranking that we found for *average* excess burdens in our earlier study (1982, Table 10).

Simple models would lead us to expect that *MEBs* would be high when activities are taxed at high or widely dispersed rates. This is borne out by our results. The labor tax rates at the industry level are fairly low and rather uniform among sectors, and the *MEBs* associated with these taxes are low. Capital tax rates are high and widely dispersed (see Table 2). Except in the case of a zero saving elasticity and a labor supply elasticity of 0.15, capital taxes have among the highest *MEBs*. We can also see the point about high and dispersed tax rates causing large *MEBs* if we look at the results for consumer sales taxes. When we raise all sales and excise taxes including the very high taxes on alcohol, tobacco, and gasoline, we have high *MEBs*. However, when we raise only the low taxes on the other commodities, we end up with very modest *MEBs*.

## V. Conclusion

In this paper we report estimates for the United States of the marginal excess burden of raising additional tax revenues. We use a dynamic sequenced numerical general equilibrium model of the U.S. economy and tax system which we have previously used to analyze specific policy proposals, such as

corporate tax integration or a move towards a consumption tax. Estimates are obtained by increasing tax rates for existing distortionary taxes.

The subject of marginal welfare costs of taxes has been discussed in the past by Harry Campbell, Browning, Dan Usher, and Charles Stuart (1984). Our contribution is in investigating this subject through a large-scale numerical general equilibrium model of the U.S. economy and tax system, incorporating all major U.S. taxes.

The central theme emerging from results is that the marginal welfare costs from raising existing distorting taxes in the United States are large, in the range of 17 to 56 cents. This has important implications for a range of policy issues. In the cost-benefit area, if a public project must be financed by distortionary taxes, the additional excess burden of these taxes should be taken into account. If this deadweight loss is as large as we suggest, it is possible that many projects accepted in recent years on the basis of favorable cost-benefit ratios should not have been undertaken. In approaching tax reform, these results suggest that a large portion of the potential welfare gains from removing distortionary taxes can be realized by a modest reduction in tax rates. Tax rate changes may, therefore, be more important than the structural reform of the tax system. In evaluating the redistribution-efficiency tradeoff in policy design, additional transfers financed at the margin by raising distorting taxes become very costly.

The issue of the marginal welfare cost of distortionary taxation has attracted increas-

ing attention during the last decade. Campbell estimated that the marginal excess burden of Canadian commodity taxes is about 24 cents. Browning reached a similar conclusion in a brief discussion of commodity taxes, but focused primarily on labor income taxes. Browning estimated that the *MEB* of these taxes is in the range of 9 to 16 cents. Browning made the conservative assumption that the compensated labor supply elasticity is 0.2. This value of the compensated elasticity is close to the values that we have when we assume that the uncompensated elasticity is zero. When we use the zero elasticity, our *MEB* estimates are only slightly higher than those of Browning. Stuart, like Browning, focuses on distortions of the labor supply decision. In his central simulations, using Browning's elasticity value, he finds *MEBs* in the range of 20.7 to 24.4 cents. Ingemar Hansson and Stuart have calculated a wide range of *MEBs* for Sweden. Their central estimates are much higher than ours or those of Browning, ranging from 69 cents to \$1.29. However, the difference can be explained by the fact that their central estimates incorporate the extremely high marginal tax rates (around 70 percent) that exist in Sweden. When Hansson and Stuart leave the rest of their model unchanged but assume marginal tax rates of 40 percent, their central case yields *MEB* estimates of from 7 to 16 cents. We feel that all of these studies point to the general conclusion that marginal excess burdens are fairly substantial. It may be too early to say that there is a consensus on this issue, but we do feel that there is growing evidence that *MEBs* may be in the range of 15 to 50 cents for an economy like that of the United States. We hope that the large estimates we report will contribute to future debate on tax reform in the United States and to a discussion of possibly modifying the cost-benefit criterion for public goods evaluation.

## REFERENCES

- Atkinson, A. B. and Stern, N. H., "Pigou, Taxation, and Public Goods," *Review of Economic Studies*, January 1974, 41, 119-28.
- \_\_\_\_\_, and Stiglitz, J. E., "The Structure of Indirect Taxation and Economic Efficiency," *Journal of Public Economics*, April 1972, 1, 97-119.
- Ballard, Charles L. and Goulder, Lawrence H., "Expectations in Numerical General Equilibrium Models," Factor Markets Workshop Research Paper No. 31, Department of Economics, Stanford University, September 1982.
- \_\_\_\_\_, Shoven, John B. and Whalley, John, "The Welfare Cost of Distortions in the United States Tax System: A General Equilibrium Approach," Working Paper No. 1043, National Bureau of Economic Research, December 1982.
- Ballard et al., Charles L., *A General Equilibrium Model for Tax Policy Evaluation*, Chicago: University of Chicago Press, 1985, forthcoming.
- Borjas, George J. and Heckman, James J., "Labor Supply Estimates for Public Policy Evaluation," *Proceedings*, Industrial Relations Research Association, 1978, 320-31.
- Boskin, Michael J., "Taxation, Saving and the Rate of Interest," *Journal of Political Economy*, April 1978, 86, S3-S27.
- Browning, Edgar K., "The Marginal Cost of Public Funds," *Journal of Political Economy*, April 1976, 84, 283-98.
- Campbell, Harry, "Deadweight Loss and Commodity Taxation in Canada," *Canadian Journal of Economics*, August 1975, 8, 441-46.
- Dasgupta, Partha S. and Stiglitz, Joseph E., "On Optimal Taxation and Public Production," *Review of Economic Studies*, January 1972, 39, 87-103.
- Denison, Edward F., "A Note on Private Saving," *Review of Economic Statistics*, August 1958, 40, 261-67.
- Diamond, Peter A. and Mirrlees, James A., (1971a) "Optimal Taxation and Public Production: I—Production Efficiency," *American Economic Review*, March 1971, 61, 8-27.
- \_\_\_\_\_, and \_\_\_\_\_, (1971b) "Optimal Taxation and Public Production: II—Tax Rules," *American Economic Review*, June 1971, 61, 261-78.
- Fullerton, Don, "Which Effective Tax Rate?," *National Tax Journal*, March 1984, 37,

23-43.

- \_\_\_\_\_, and Henderson, Yolanda K., "Incentive Effects of Taxes on Income from Capital: Alternative Policies in the 1980's," Discussion Paper No. 61, Woodrow Wilson School, Princeton University, December 1983.
- \_\_\_\_\_, Shoven, John B. and Whalley, John, "Replacing the U.S. Income Tax With a Progressive Consumption Tax: A Sequenced General Equilibrium Approach," *Journal of Public Economics*, February 1983, 20, 3-23.
- Fullerton, et al., Don, "Corporate Tax Integration in the United States: A General Equilibrium Approach," *American Economic Review*, September 1981, 71, 677-91.
- Hansson, Ingemar and Stuart, Charles, "Tax Revenue and the Marginal Cost of Public Funds in Sweden," mimeo., University of California-Santa Barbara, January 1983.
- Harberger, Arnold C., "The Incidence of the Corporation Income Tax," *Journal of Political Economy*, June 1962, 70, 215-40.
- \_\_\_\_\_, "Efficiency Effects of Taxes on Income from Capital," in Marian Krzyzaniak, ed., *Effects of Corporation Income Tax*, Detroit: Wayne State University Press, 1966, ch. 15.
- Hausman, Jerry, "Labor Supply," in Henry J. Aaron and Joseph A. Pechman, eds., *How Taxes Affect Economic Behavior*, Washington: The Brookings Institution, 1981.
- Killingsworth, Mark R., *Labor Supply*, New York: Cambridge University Press, 1982.
- Kimbell, Larry J. and Harrison, Glenn W., "General Equilibrium Analysis of Regional Fiscal Incidence," in Herbert Scarf and John B. Shoven, eds., *Applied General Equilibrium Analysis*, New York: Cambridge University Press, 1984, ch. 7.
- Menchik, Paul L. and David, Martin, "Income Distribution, Lifetime Savings, and Bequests," *American Economic Review*, September 1983, 73, 672-90.
- Merrill, O. H., "Applications and Extensions of an Algorithm that Computes Fixed Points to Certain Upper Semi-Continuous Point-to-Set Mappings," unpublished doctoral dissertation, University of Michigan, 1972.
- Musgrave, Richard A., *The Theory of Public Finance*, New York: McGraw-Hill, 1959.
- Samuelson, Paul A., "The Pure Theory of Public Expenditure," *Review of Economics and Statistics*, November 1954, 36, 387-89.
- Starrett, David A., "Long Run Savings Elasticities in the Life Cycle Model," Factor Markets Workshop Research Paper No. 24, Stanford University, August 1982.
- Stiglitz, Joseph E., "Taxation, Corporate Financial Policy, and the Cost of Capital," *Journal of Public Economics*, February 1973, 2, 1-34.
- Stuart, Charles E., "Welfare Costs per Dollar of Additional Tax Revenue in the United States," *American Economic Review*, June 1984, 74, 352-62.
- Summers, Lawrence H., "Capital Taxation and Accumulation in a Life Cycle Growth Model," *American Economic Review*, September 1981, 71, 533-44.
- Usher, Dan "The Private Cost of Public Funds: Variations on Themes by Browning, Atkinson, and Stern," Discussion Paper No. 481, Institute for Economic Research, Queen's University, June 1982.
- Board of Governors of the Federal Reserve System, *Flow of Funds Accounts, 1946-75*, Washington, 1976.
- U.S. Department of Commerce, Bureau of Economic Analysis, "U.S. National Income and Products Accounts, 1973 to Second Quarter 1976," *Survey of Current Business*, July 1976, 56, 22-69.
- \_\_\_\_\_, *The Detailed Input-Output Structure of the U.S. Economy: 1972*, Washington: USGPO, 1979.

# The Use of Protection and Subsidies for Entry Promotion and Deterrence

By AVINASH K. DIXIT AND ALBERT S. KYLE\*

Artificial barriers to international trade and natural barriers to entry are both important features of certain imperfectly competitive high-technology industries, such as aerospace and computers. The effort by the European Airbus consortium to compete with Boeing in the market for intermediate-range commercial jets is the best known instance. It has been argued that protection of the Airbus' home market, and subsidies from the partner governments, allow the consortium to recover a significant fraction of the huge sunk costs of developing the Airbus. It has also been argued that without such support, and perhaps, also without access to Boeing's home market, the entire project would not be viable. In the United States, the protection and subsidies are regarded as unfair practices, requiring countermeasures.<sup>1</sup> A similar argument has been made about Japan's entry into the markets for 16K and 64K RAM microchips. The boot may be on the other foot for the 256K and megabit generations, with Japan dominant and the United States attempting entry.<sup>2</sup>

The aim of this paper is to begin analysis of the functioning of such markets and the role of policies towards them. To model the issues adequately, the potential for strategic behavior on part of both governments and firms must be taken into the account. Furthermore, it is important to recognize that the strategies of governments interact with

those of firms. The appropriate model is therefore a game-theoretic one, with the governments and the firms as the players.

The games we study differ in the order in which the players move. The issue is not merely, or even primarily, one of timing. It is whether one player can make a strategic precommitment of policy choice that affects subsequent choices of others to his own advantage. The availability of such moves in practice will depend on particular features of governmental and international institutions. Here we merely make different assumptions in this regard and compare the outcomes, thus determining the effects that particular strategies would have when available.

All the games have one common feature. The players who move earlier are aware of the nature of the game to follow, and correctly forecast the outcomes of those stages when computing the consequences of their own current actions. Therefore the games are solved backwards: the optimal actions at the last stage are calculated for all conceivable patterns of previous actions, these are used in determining the optimal choices at all penultimate-stage situations that can arise, and so on. The solution of the game as a whole has the property of rational expectations, and is commonly called a (subgame) perfect equilibrium.<sup>3</sup>

The particular theoretical model, and its notation, are suggested by the Airbus example. This should not be taken literally; we make several restrictive assumptions, so as to highlight and illustrate the strategic aspects in the simplest possible format. In the concluding section we will comment on the consequences of relaxing some of these assumptions.

\*Woodrow Wilson School, Princeton University, Princeton, NJ 08544. Dixit gratefully acknowledges support from the National Science Foundation under grant SES-8308536. Kyle thanks the Centre of Policy Studies, Monash University, Australia for its hospitality and support when this research was done. We are both grateful to Peter Hartley for many helpful discussions, and to Gene Grossman and two referees for comments on an earlier version.

<sup>1</sup>*The Economist*, August 27, 1983, pp. 12-13.

<sup>2</sup>*The Economist*, June 19, 1982, Survey, p. 14.

<sup>3</sup>See Reinhard Selten (1975). A simplified exposition adequate for our concerns can be found in Dixit (1982).

We assume that there are only two countries and two firms. One of the firms is an incumbent (i.e., its sunk costs have already been incurred), located in one of the countries (*US*). The other firm is a potential entrant located in the other country (*EC*). The analysis is partial equilibrium in nature, that is, income effects are ignored and factor prices are exogenous. The objective of each firm is to maximize its profit. The objective of each government is taken to be the standard criterion of social welfare in partial equilibrium ignoring distribution, that is, the sum of *domestic* consumers' and producers' surpluses.

To simplify further, we restrict the policy choices. Each government has an all-or-nothing trade policy choice: free trade or complete prohibition of all imports in the industry. Similarly, when the *EC* government uses a subsidy policy, the choice is one of financing all or none of the sunk costs of its firm. The advantage of these restrictions is that we can use the "reduced-form" outcomes of the price and output decisions made by the monopolists or duopolists in the equilibrium consequent upon the discrete policy choices. The binary comparisons of profits and surpluses then govern the decisions to be made. Thus we are not tied down to any specific assumption such as Cournot or Bertrand behavior for the duopoly.<sup>4</sup>

The operation of the model, and the results, are in general conformity with economic intuition drawn from both industrial organization and international trade. The inability of firms to appropriate all of consumers' surplus can justify a policy to alter an oligopolistic market outcome even from the viewpoint of worldwide efficiency.<sup>5</sup> For a single country, the capture of any monopoly rents on behalf of its own residents can make it desirable to pursue policies that favor its

own firms and harm foreign ones.<sup>6</sup> When other governments are simultaneously pursuing trade policies, a prisoner's dilemma can arise at the policy level.<sup>7</sup> In our model, all these possibilities exist, but they take on new forms because of the irreversibility of entry. The timing of policy actions, and the degree of commitment to them, become crucial. Therefore these are the aspects we emphasize.

In Section I, the model is specified in more detail, and in Sections II–V its implications for strategic uses of policies towards entry are examined. We see how and when such policies promote national advantage. We find that the *EC* government gains from protectionist entry-promotion whenever entry occurs with such a policy but not without it. We also examine the implications for world welfare. As a general tendency, protection for entry promotion is harmful to the world, and countermeasures that deter such a policy are beneficial. Subsidies as instruments of entry promotion are generally more desirable from a world viewpoint; countermeasures against them are either ineffective or harmful.

In Section VI we indicate how the approach developed here can be adapted to extend the model by changing some of the special assumptions.

## I. Structure and Notation

Consider a homogeneous good that can be produced by a firm after it has undertaken a preproduction (sunk) expenditure  $K$ , at constant unit production cost  $c$ . International cost differences are ignored so as to focus on strategic interactions. There are two markets, *US* and *EC*, each with its own demand curve, and no resale by third parties. There is an incumbent *US* firm and a potential *EC* entrant firm.

<sup>4</sup>For a different approach, see Douglas Curtis (1983). He considers entry promotion with particular solution concepts, but does not consider strategic behavior by the incumbent firm or government.

<sup>5</sup>See Michael Spence (1976), and Dixit and Joseph Stiglitz (1977).

<sup>6</sup>See James Brander and Barbara Spencer (1984), Jonathan Eaton and Gene Grossman (1983), and surveys in Dixit (1984), and Grossman and David Richardson (1984).

<sup>7</sup>See Harry Johnson (1954).



In the *US* market, let  $PU_1$  = excess of revenue over production cost for a monopolist,  $PU_2$  = excess of revenue over production cost for each duopolist,  $SU_1$  = consumers' surplus under monopoly, and  $SU_2$  = consumers' surplus under duopoly. Consumers' surplus is measured with reference to some fixed price higher than any which arises in the market. For a nontrivial problem the demand is positive at price  $c$ . Then  $PU_1 > 0$  and  $SU_1 > 0$ . Also  $PU_2 \geq 0$  since zero production is always feasible. More importantly,

$$(1) \quad PU_1 \geq 2PU_2,$$

since a monopolist can always mimic the actions of two duopolists, and

$$(2) \quad 2PU_2 + SU_2 > PU_1 + SU_1,$$

assuming closed-economy duopoly is socially preferable to monopoly (because of higher output and lower price). This holds for a wide range of duopoly equilibrium concepts, including the familiar Bertrand and Cournot ones, so we do not need to specify any particular duopoly model. In Bertrand competition, for example, price equals production cost  $c$ , so  $PU_2 = 0$ .

In the *EC* market, we define  $PE_1$ ,  $PE_2$ ,  $SE_1$ , and  $SE_2$  in an analogous manner; require  $PE_1 > 0$ ,  $SE_1 > 0$ ,  $PE_2 \geq 0$ ; and have

$$(3) \quad PE_1 \geq 2PE_2,$$

$$(4) \quad 2PE_2 + SE_2 > PE_1 + SE_1.$$

Equilibrium is the outcome of a game played by three players: the *US* government, the *EC* government, and the potential entrant *EC* firm.<sup>8</sup> The two governments choose as their actions either free trade or protection, where "protection" is defined as the complete prohibition of imports. The *EC*

government decides whether to subsidize its firm's sunk costs. The *EC* firm decides whether to enter. The eight combinations of trade policy and entry decisions generate six different possible outcomes because *US* trade policy does not affect the outcome when the *EC* firm chooses not to enter. The six outcomes are the following:

*Autarky (AUT)*: The *EC* firm enters and both governments protect. Each market is served by its own monopoly firm.

*Protected Entry (PRE)*: The *EC* firm enters, the *EC* government protects, and the *US* government has free trade. There is duopoly in the *US* market and a monopoly by the *EC* firm in the *EC* market.

*Free Trade Duopoly (FTD)*: The *EC* firm enters and both governments choose free trade. There is duopoly in both markets.

*US Monopoly (MON)*: The *EC* firm does not enter, the *EC* government has free trade, and *US* government policy does not matter. The *US* firm has a monopoly in both markets.

*US Protection (USP)*: The *EC* firm enters, the *US* protects, and the *EC* has free trade. The *US* firm has a monopoly in the *US* market, and there is duopoly in the *EC* market.

*Irrational Protection (IRR)*: The *EC* firm does not enter, the *EC* protects, and *US* trade policy does not matter. There is monopoly in the *US* market, and the *EC* market is not served at all.

The objective of each government is to maximize the sum of domestic firm profits and domestic consumers' surplus. The objective of the *EC* firm is to maximize profits (net of sunk costs).

The value of the six outcomes to the three players is summarized in Table 1. Of these six outcomes, it is immediately obvious that the sixth outcome, "irrational protection," never arises as a perfect equilibrium outcome of the game. By choosing free trade when

<sup>8</sup>The *US* firm is a player in the "background" game of duopoly that is played if entry occurs. But it has already sunk its  $K$  and has no strategic choices, so it is not modeled explicitly.

TABLE 1—VALUES OF OUTCOMES TO THE THREE PLAYERS

Outcome	Player		
	US Government	EC Government	EC Firm
Autarky ( <i>AUT</i> )	$SU_1 + PU_1$	$SE_1 + PE_1 - K$	$PE_1 - K$
Protected Entry ( <i>PRE</i> )	$SU_2 + PU_2$	$SE_1 + PE_1 + PU_2 - K$	$PE_1 + PU_2 - K$
Free Trade Duopoly ( <i>FTD</i> )	$SU_2 + PU_2 + PE_2$	$SE_2 + PE_2 + PU_2 - K$	$PE_2 + PU_2 - K$
US Monopoly ( <i>MON</i> )	$SU_1 + PU_1 + PE_1$	$SE_1$	0
US Protection ( <i>USP</i> )	$SU_1 + PU_1 + PE_2$	$SE_2 + PE_2 - K$	$PE_2 - K$
Irrational Protection ( <i>IRR</i> )	$SU_1 + PU_1$	0	0

protection would not lead to entry by the *EC* firm, the *EC* government achieves (at worst) the superior outcome, "US monopoly."

In strategically interesting cases, the outcome of the game depends upon how the three players order the remaining five outcomes. For each player, several orderings consistent with Table 1 are possible.

## II. Worldwide Optimality

Worldwide social surplus, defined as the sum of world consumers' surplus and world producers' profits, is measured by adding together the objectives the *US* government and the *EC* government. Using this measure of social optimality, several rankings of the outcomes in Table 1 are possible, but these rankings do satisfy certain constraints. Because domestic duopoly is better than domestic monopoly, free trade duopoly is preferred to protected entry (by equation (4)) and to *US* protection (by equation (2)), and both of these outcomes are preferred to autarky. Furthermore, *US* monopoly is preferred to autarky since the *EC* firm's sunk costs do not have to be incurred, and to irrational protection since a monopoly in the *EC* market is better than no market at all. Thus, the socially best outcome is either free trade duopoly (if entry is desirable) or *US* monopoly (if entry is not desirable). The former is optimal provided

$$(5) \quad 2PU_2 + SU_2 + 2PE_2 + SE_2 - K > PU_1 + SU_1 + PE_1 + SE_1,$$

and the latter otherwise. Neither involves any protection, but the attainment of *FTD*

can require a jointly financed subsidy to induce entry.<sup>9</sup>

Even this optimality is second best, arising from an inability of the governments to bargain efficiently with the *US* monopolist in the first place. If efficient bargaining were allowed, both governments would "bribe" the *US* monopolist to supply the quantity demanded at marginal cost *c* in their respective domestic markets. Entry would never occur.

Inequality (5) is more likely to be satisfied (i) the more competitive the duopoly, and (ii) the lower the sunk cost. In any actual instance, these can work in opposite ways. In the aircraft industry, for example, the competition between Boeing and Airbus is fierce, but sunk costs are large. Considering the size of the world market, however, our presumption would be that in most industries (5) would hold, that is, *FTD* would be better than *MON*.

Later we will come across similar comparisons. From a world viewpoint, *PRE* is preferred to *MON* if

$$(6) \quad 2PU_2 + SU_2 - K > PU_1 + SU_1,$$

and *USP* is preferred to *MON* if

$$(7) \quad 2PE_2 + SE_2 - K > PE_1 + SE_1.$$

In each of these, the sunk cost is compared to the gain in total surplus in only one of the

<sup>9</sup>This assumes that the subsidy is financed with non-distorting taxes. Otherwise, (5) must be modified by subtracting the deadweight loss associated with the taxes from the left-hand side.

markets. Thus (6) and (7) are less likely to be satisfied than is (5). For a high sunk cost industry like aircraft, it seems likely that (6) and (7) would not hold.

### III. Nonstrategic Trade Policy Choices

We provide a reference point for the analysis of strategic trade policy choices by first considering a game in which they have no scope. This is done by placing the firm's entry decision first, and the governments' policies later. Such a game will be played if the governments are in fact unable to make credible policy commitments. We number this as game 1, and call it the reference game.

As explained above, the second stage is solved first. If the firm stays out, the *EC* government will choose free trade, the *US* policy will be immaterial, and the outcome will be *MON*. If the firm enters, a subsidy is irrelevant, but the trade policies of the governments interact as shown in Table 2.

Reference to Table 1 shows that each government has a dominant strategy (and therefore the order of their moves is immaterial). For example, the *US* government prefers protection if

$$(8) \quad PU_1 + SU_1 > PU_2 + SU_2,$$

and free trade otherwise. When the *EC* policy is protection, these are just the *US* payoffs from the respective choices. When the *EC* policy is free trade, the duopoly profit in that market,  $PE_2$ , is added to each side to get the *US* payoffs, and the comparison is unaffected. Note how our assumptions of constant marginal costs, and no international resales, separate the markets and yield this result. Similarly, the *EC* government's dominant policy is protection if

$$(9) \quad PE_1 + SE_1 > PE_2 + SE_2,$$

and free trade otherwise. Since these choices are determined by the sum of consumers' surplus and domestic firm profits in the domestic market alone, we call them the *domestically preferred trade policies*. There are

TABLE 2—POST-ENTRY TRADE POLICY GAME

US	EC	
	Protection	Free Trade
Protection	<i>AUT</i>	<i>USP</i>
Free Trade	<i>PRE</i>	<i>FTD</i>

four possible combinations. The two countries may have different domestically preferred policies either because duopoly solution concepts differ (on account of, say, different antitrust policies), or because the demand curves in the two countries have different shapes. At any rate, it is necessary to consider all four combinations of domestically preferred policies, even though only the symmetric ones will be relevant when the two markets are identical.

We mention in passing the possibilities for some commonly used solution concepts. If the duopoly is collusive (price and quantities same as in monopoly), the domestically preferred policy is protection. If the duopoly is Bertrand (price equals  $c$ ), it is free trade. For a Cournot duopoly, routine calculations show that the domestically preferred policy is protection with a linear demand curve ( $p = a - bq$ ,  $a > c$ ), and free trade with an exponential demand curve ( $p = ae^{-bx}$ ,  $c = 0$ ).

Now the *EC* firm knows that, if it enters, each government will choose its domestically preferred policy. Therefore it will assume these policies to test whether entry yields positive profit (net of sunk costs). This depends on the size of  $K$ . We therefore have another dimension of classification, namely the necessary and sufficient conditions for profitable entry. The four interesting categories follow:

- (10) Protection (of the *EC* market) and Access (to the *US* market):

$$\max(PE_1, PE_2 + PU_2) < K < PE_1 + PU_2;$$

- (11) Protection:  $PE_2 + PU_2 < K < PE_1;$

- (12) Access:  $PE_1 < K < PE_2 + PU_2;$

TABLE 3—OUTCOMES OF TRADE POLICY GAMES

Conditions for Profitable Entry	Domestically Preferred Policies			
	US Protection EC Protection	US Free Trade EC Protection	US Protection EC Free Trade	US Free Trade EC Free Trade
Protection and Access	1. <i>MON</i> 2. <i>MON</i> 3. <i>MON</i> 4. <i>MON</i>	1. <i>PRE</i> 2. <i>PRE</i> 3. $\text{Max}_{US}(\text{MON}, \text{PRE})^c$ 4. $\text{Max}_{US}(\text{MON}, \text{PRE})$	1. <i>MON</i> 2. <i>MON</i> 3. <i>MON</i> 4. <i>MON</i>	1. <i>MON</i> 2. <i>PRE</i> <sup>a</sup> 3. $\text{Max}_{US}(\text{MON}, \text{PRE})^c$ 4. $\text{Max}_{US}(\text{MON}, \text{PRE})$
Protection	1. <i>AUT</i> 2. <i>AUT</i> 3. <i>AUT</i> 4. <i>AUT</i>	1. <i>PRE</i> 2. <i>PRE</i> 3. <i>PRE</i> 4. <i>PRE</i>	1. <i>MON</i> 2. <i>AUT</i> <sup>a</sup> 3. <i>AUT</i> 4. <i>AUT</i>	1. <i>MON</i> 2. <i>PRE</i> <sup>a</sup> 3. <i>PRE</i> 4. <i>PRE</i>
Access	1. <i>MON</i> 2. <i>MON</i> 3. <i>MON</i> 4. <i>MON</i>	1. <i>PRE</i> 2. <i>PRE</i> 3. $\text{Max}_{US}(\text{MON}, \text{PRE})^c$ 4. $\text{Max}_{US}(\text{MON}, \text{FTD})^e$	1. <i>MON</i> 2. <i>MON</i> 3. <i>MON</i> 4. <i>MON</i>	1. <i>FTD</i> 2. <i>FTD</i> <sup>b</sup> 3. $\text{Max}_{US}(\text{MON}, \text{FTD})^c$ 4. $\text{Max}_{US}(\text{MON}, \text{FTD})$
Protection or Access	1. <i>AUT</i> 2. <i>AUT</i> 3. <i>AUT</i> 4. $\text{Max}_{US}(\text{AUT}, \text{Max}_{EC}(\text{AUT}, \text{FTD}))^e$	1. <i>PRE</i> 2. <i>PRE</i> 3. <i>PRE</i> 4. $\text{Max}_{US}(\text{PRE}, \text{Max}_{EC}(\text{AUT}, \text{FTD}))^f$	1. <i>MON</i> 2. <i>AUT</i> <sup>a</sup> 3. $\text{Max}_{US}(\text{AUT}, \text{FTD})^d$ 4. $\text{Max}_{US}(\text{AUT}, \text{FTD})$	1. <i>FTD</i> 2. <i>FTD</i> <sup>b</sup> 3. <i>FTD</i> 4. <i>FTD</i>

Note: The order of play:

1. Reference Game: The EC firm makes its entry decision first; the governments choose their (domestically preferred) policies later.

2. The Entry Promotion Game: The EC government chooses its trade policy first, then the EC firm makes its entry decision, and finally the US government chooses its (domestically preferred) policy.

3. Entry Deterrence by Commitment: The US government chooses its trade policy first, the EC government next, and the firm's entry decision is last.

4. Threats and Promises: The US government chooses a policy rule first, but it can specify a policy choice conditional on the policy choice made by the EC government at the second stage. The EC firm's entry decision is last.

<sup>a</sup>These outcomes are instances of entry promotion through protection.

<sup>b</sup>If the order of play of the EC firm and the US government is reversed, the outcome changes to *PRE* if the US government prefers *MON* to *FTD*; in all other cells this change of order does not affect the outcome.

<sup>c</sup>When the outcome is *MON*, the US policy is one of "entry deterrence."

<sup>d</sup>When the outcome is *FTD*, the US policy is one of "dissuaded entry promotion."

<sup>e</sup>When the outcome is *FTD*, it results from an "optimal promise" by the US government.

<sup>f</sup>When the outcome is *FTD*, it results from an "optimal threat" by the US government.

<sup>g</sup>When the outcome is *FTD*, it results from (i) an "optimal threat" if the US government prefers *PRE* to *MON* (i.e., the outcome of game 3 is *PRE*), and (ii) an "optimal promise" if the US government prefers *MON* to *PRE* (i.e., the outcome of game 3 is *MON*).

### (13) Protection or Access:

$$PE_2 < K < \min(PE_1, PE_2 + PU_2).$$

Note that the separate cases of Protection and Access are mutually exclusive, that is, only one arises for each given set of values of  $PE_1$ ,  $PE_2$ , and  $PU_2$ . We omit the categories where neither protection nor access is needed

( $K < PE_2$ ), and where both are not sufficient ( $K > PE_1 + PU_2$ ), since the role of strategic trade policies is not of interest in these cases. The latter will become relevant in Section V, where subsidies can be employed to promote unprofitable entry if the increase in consumers' surplus is large enough.

Considering the domestically preferred policies and the conditions of entry together, there are sixteen cases. We provide a table of

outcomes for completeness. But to avoid excessive taxonomy, we focus attention on a few cases of special interest, and then state some general results.<sup>10</sup>

The outcomes of games involving trade policies are shown in Table 3. Recall that the reference game has number 1. In the last column, for example, both domestically preferred policies are free trade. Therefore entry occurs when access to the *US* market suffices for profitability (Access, and Protection or Access) and the outcome is *FTD*. There is no entry in the other two cases, and the outcome is *MON*.

#### IV. Strategic Use of Trade Policies

Now we modify the reference game by allowing one of the governments to move first and make an irreversible, precommitted, policy choice. We then examine how such policies can be used for entry promotion by the *EC* government and deterrence by the *US* government. This section considers trade policies; the role of subsidies is taken up in the next section.

Precommitment yields new outcomes only to the extent that the chosen policy differs from the domestically preferred one. We shall see how such a commitment can be desirable to the government making it. But its credibility is problematic, since after the entry decision is made, the government has the incentive to reverse its choice. In practice there are devices that make such reversals costly; examples are constitutional provisions, automatic administrative procedures, the power of the special interests that favor the committed policy, and the government's reputation. Their reliability is a matter of degree. The question of which government is better placed to seize the initiative and make a commitment is also open. Theory at this level of generality can only show the consequences of the alternative possibilities in these matters, and that is the approach we take.

We consider three games of this kind. The first concerns entry promotion, and is numbered 2 in the table of outcomes (the reference game being number 1). Here the *EC* government moves first and chooses its trade policy. The *EC* firm moves second and decides whether to enter. The *US* government moves last and chooses its (domestically preferred) policy. (Reversing the last two stages would produce a different game, but in all cases its outcomes are Pareto worse, i.e., worse for all players, or the same. Thus, given that the *EC* government moves first, the *US* government would choose to move third rather than second. Therefore the order we have assumed is appropriate.)

Game 3 considers entry deterrence. The *US* government moves first, and chooses protection or free trade. The *EC* government moves second to choose its trade policy, and the *EC* firm's entry decision is last. (Reversing the last two stages would not be in the interests of either the *EC* government or the *EC* firm.)

Game 4 is similar, but the *US* government at the first move has an additional choice. It can precommit to a policy rule, which specifies a choice of protection or free trade conditional on the action taken by the *EC* government at its move. Such conditional rules can be of two kinds, "matching," defined as protection when the *EC* policy is protection and free trade when the *EC* policy is free trade, and "reversing," defined as protection when the *EC* policy is free trade and free trade when it is protection. The reversing policy is, however, never optimal because each government prefers that the other choose free trade and a reversing policy thus encourages the *EC* government to be protectionist which the *US* government does not want. The outcome of the conditional commitment game, if different from the unconditional one, must therefore be the result of a matching policy. Depending on circumstances, this can act as a threat (we will protect our market if you protect yours), or a promise (we will practice free trade if you do). In reality, such policies can be found in the form of reciprocity legislation or procedures, but their credibility is again open to

<sup>10</sup> Details of the arguments concerning trade policies (summarized in Table 3) can be found in our earlier paper (1983).

doubt. The earlier remarks on this issue apply here, too.

It is clear from Table 3 that the outcomes of games 2, 3, and 4 are in many cases different from those of the reference game. These different outcomes result from strategic policies which illustrate various strategic themes: entry promotion, entry deterrence, dissuasion, threats and promises. In the rest of this section we discuss these themes, illustrate them with specific cells from Table 3, and point out general tendencies for world welfare. Here we will not compute perfect equilibria in detail from the game trees. The verbal arguments are usually evident enough, and the interested reader can easily construct the details.

#### A. Entry Promotion

In the reference game, *MON* is in some cases the outcome even though a shift in *EC* policy from free trade to protection (holding *US* policy constant) makes entry profitable. In game 2, the *EC* can commit to protection and thus promote entry. In fact, the *EC* government will always make such a commitment in these cases. The simple explanation is that the *EC* market is going to be monopolized anyway, by the *US* firm if entry does not occur and by the protected *EC* firm if entry does occur. Since no gain or loss in consumers' surplus is at stake, the only question is whether a protected *EC* firm can make a profit, as a domestic monopolist if the *US* market is closed, or as both a domestic monopolist *cum* exporting duopolist if the *US* market is open.

Entry promotion shifts the outcome from *MON* to *AUT* when the *US* policy is protection, and from *MON* to *PRE* when the *US* policy is free trade. While the shift from *MON* to *AUT* always lowers world surplus, a shift from *MON* to *PRE* raises world surplus if inequality (6) is satisfied. Thus we cannot state as a universal rule that strategic entry promotion lowers world welfare. There is some presumption that it does so when sunk costs are large relative to the *US* market alone or when a *US* duopoly is not very

different from a monopoly (because, say, collusion is easy).

*Example:* Consider the case where both governments' domestically preferred policy is free trade, and profitable entry by the *EC* firm requires both protection of the *EC* market and access to the *US* market. In obvious notation, this is cell (1,4) in Table 3. This case can arise when the duopoly is very competitive and the sunk costs are large, and is therefore of special interest for the aircraft industry. In the reference game, the firm knew that protection would not be provided after entry. It kept out, and *MON* was the outcome. Now suppose the *EC* government commits itself to a protectionist policy. Knowing that the *US* government's *ex post* choice is free trade, the *EC* firm will expect positive profits and therefore will enter. The outcome will be *PRE*. The *EC* government prefers this to *MON*; the additional payoff ( $PE_1 + PU_2 - K$ ) is exactly the *EC* firm's profit. Therefore, the *EC* government will make the commitment in game 2.

#### B. Entry Deterrence

In games 1 and 2, there are cases where entry occurs but a shift in *US* policy from free trade to protection (holding *EC* policy fixed) makes entry unprofitable. Game 3 has a structure allowing the *US* government to commit to protection and thus deter entry in these cases. It is not always optimal for the *US* government to make such a commitment. It does so if and only if the *MON* achieved by entry deterrence is better for the *US* than the *FTD* or *PRE* which occurs in games 1 or 2. Note that *MON* is preferred by the *US* whenever the producer gains from eliminating a competitor at home and perhaps also abroad outweigh the consumers' surplus lost in the *US* market.

From the point of view of world welfare, there is some presumption that aggressive entry deterrence which shifts the outcome from *FTD* to *MON* is undesirable, as discussed in Section III. A shift from *PRE* to *MON* is desirable provided inequality (6) does not hold. Certainly, whenever entry

promotion is bad, entry deterrence which reverses the outcome is good.

*Example:* (Entry deterrence which counters entry promotion): Return to cell (1,4) and consider game 3. If the *US* government commits itself to protection, then any later *EC* moves to promote entry through protection will fail for want of access to the *US* market. Therefore such moves will not be made and the *US* commitment will produce *MON*. The *US* prefers this to *PRE* and makes the commitment when

$$SU_1 + PU_1 + PE_1 > SU_2 + PU_2,$$

that is, the monopoly profits in the *EC* market outweigh the *US* domestic preference for free trade.

*Example:* (Aggressive entry deterrence): Consider cell (3,4), game 3. By denying access the *US* government can aggressively prevent unassisted entry and achieve *MON* rather than *FTD*, if the *EC* market offers sufficiently higher profits. Our presumption is that such conduct lowers world welfare.

#### C. Dissuasion

It is possible that a commitment by the *US* government to free trade can forestall entry promotion by making it unnecessary for the *EC* to use protection to promote entry. Game 3 has a structure allowing such a commitment, which we refer to as "dissuasion." Since dissuasion always converts *AUT* to *FTD*, it clearly increases world surplus. In fact, it makes both governments better off, as the following example—the only case of dissuasion in Table 3—shows:

*Example:* Consider cell (4,3). Since access in game 2 is unavailable given the *US* domestic preference for protection, the *EC* government goes against its domestic preference and promotes entry via protection. The *US* can avoid this in game 3 by going against its domestic preference and committing to free trade. It prefers the resulting *FTD* to

*AUT* if

$$SU_2 + PU_2 + PE_2 > SU_1 + PU_1,$$

that is, the duopoly profits in the *EC* market are large enough to override the *US* domestic preference. Then it will provide the commitment and the result will be good for both governments.

#### D. Threats and Promises

When a matching strategy in game 4 results in an outcome different from game 3, the changed outcome is the result of a threat or a promise. In a promise, the *US* government changes its policy from protection to free trade in order to get the *EC* government to change its policy as well. In the actual outcome from a threat, the *US* policy does not change (because it is already free trade) but the *EC* is induced to change its policy from protection to free trade. Clearly, threats and promises always lead to *FTD* when they change the outcome of game 3. While the *US* is better off as a result of the availability of threats and promises, the *EC* is made better off by promises, worse off by threats. Both threats and promises, however, always increase world surplus.

*Example (Promise):* Consider cell (4,1). Both governments have a domestic preference for protection, and this suffices for the *EC* firm's entry, producing *AUT*. The *US* would not want to commit unconditionally to free trade; that would only produce *PRE*. But suppose it makes a conditional commitment promising to reciprocate an *EC* choice of free trade. Now the *EC* has another route to entry, namely mutual free trade that provides access. If the *EC* prefers *FTD* to *AUT*, it will pursue this. If, in turn, the *US* prefers *FTD* to *AUT*, it will make such a promise. Such joint preference for *FTD* over *AUT* is possible despite the domestic preferences for protection, if the duopoly profits in each other's market are large enough. Here, a conditional policy overcomes a prisoner's dilemma, allowing the noncooperative outcome



to be replaced by a cooperative outcome preferred by both governments.

*Example (Threat):* Consider cell (4,2). A *US* threat of protecting its own market in response to *EC* protection can lead the *EC* to opt for *FTD* over *AUT*. If the *US* prefers this to the outcome *PRE* of following its domestic preference, it will utilize such a threat. The *EC* is worse off, but world surplus is increased because an additional market is served by both firms.

The above examples illustrate that strategic trade policies work in a variety of ways when applied in different circumstances. Do the various cases suggest any per se rules that a body like GATT would want to adopt concerning such policies? Certainly threats, promises, and dissuasion—all of which lead to *FTD*—should be allowed, and entry promotion which leads to *AUT* should be disallowed. Under the presumption that entry is desirable if and only if all markets are served, a rule mandating bilateral free trade achieves the best outcome possible without subsidies or bargains with the monopolist. Such a rule eliminates all entry promotion as well as aggressive entry deterrence. In the absence of such a presumption, however, no obvious rule emerges.

## V. Strategic Use of Subsidies

Now we allow the *EC* government the additional policy instrument of paying the *EC* firm's sunk cost  $K$ . This increases the number of *EC* policy choices from two to four: "subsidized protection," "unsubsidized protection," "subsidized free trade," and "unsubsidized free trade." If the *EC* chooses to subsidize, the firm will enter irrespective of trade policies.<sup>11</sup> The *US* government can choose its trade policy to affect the *EC* decision in this respect. Countervailing duties are an instance. Both governments also use trade

policies for familiar reasons of domestic preference.

Define games 5–7 as the respective modifications of games 2–4. In game 5, the *EC* government moves first to choose its trade and subsidy policies. In game 6 the *US* government moves first to choose its trade policy. In game 7 it chooses a policy rule contingent on the second stage trade and subsidy choices of the *EC* government. "Matching" is now defined as free trade if the *EC* government chooses unsubsidized free trade, and protection otherwise.

The most interesting cases to consider are those where the *EC* government's domestically preferred policy is free trade. Table 4 shows these cases, which correspond to the last two columns of Table 3. A fifth row has been added to handle the case where sunk costs are so high that entry is unprofitable regardless of trade policies. This case is uninteresting in the absence of subsidies, but must now be considered. Game 7 is not shown separately because its outcome in each of these cases is the same as that of game 6, that is, conditional policies do not give the *US* any additional leverage when the *EC* is able to subsidize.

### A. Entry Promotion

Now use Tables 3 and 4 to compare the outcomes of games 5 and 2. Because the *EC* government's domestically preferred policy is free trade, subsidized free trade is a better instrument for entry promotion than unsubsidized protection. This results in two kinds of changes in outcomes, depending upon whether entry occurs in game 2. When entry does occur in game 2, the *EC* government changes its policy to subsidized free trade in game 5 and thus improves the outcome for both countries. The *EC* is better off because its domestically preferred policy is free trade, the *US* because the *EC* market is opened. The outcome changes from *PRE* to *FTD* or from *AUT* to *USP*, depending upon whether the *US* market is open or closed. In cases where entry does not occur in game 2, the *EC* government sometimes promotes entry with subsidized free trade in game 5. This shifts the outcome from *MON* to *USP* or

<sup>11</sup>The binary choice of paying  $K$  or 0 does not impose any restrictions since any payment in excess of what is needed to induce entry, or short of this level, is a transfer that cancels out in the *EC* government's payoff.



TABLE 4—OUTCOMES OF SUBSIDY GAMES

Conditions for Profitable Entry	Domestically Preferred Policies	
	US Protection EC Free Trade	US Free Trade EC Free Trade
Protection and Access	5. $\text{Max}_{EC}(MON, USP)$ 6. $\text{Max}_{EC}(MON, USP)$	5. $FTD$ 6. $\text{Max}_{US}(FTD, \text{Max}_{EC}(MON, USP))$
Protection	5. $USP$ 6. $USP$	5. $FTD$ 6. $FTD$
Access	5. $\text{Max}_{EC}(MON, USP)$ 6. $\text{Max}_{EC}(MON, USP)$	5. $FTD$ 6. $\text{Max}_{US}(FTD, \text{Max}_{EC}(MON, USP))$
Protection or Access	5. $USP$ 6. $USP$	5. $FTD$ 6. $FTD$
Entry Never Profitable	5. $\text{Max}_{EC}(MON, USP)$ 6. $\text{Max}_{EC}(MON, USP)$	5. $\text{Max}_{EC}(MON, FTD)$ 6. $\text{Max}_{US}(\text{Max}_{EC}(MON, FTD), \text{Max}_{EC}(MON, USP))$

Note: The order of play:

5. Entry Promotion: The EC government chooses its trade and subsidy policies first, then the EC firm makes its entry decision, and finally the US government chooses its (domestically preferred) trade policy.

6. Entry Deterrence: The US government chooses its trade policy (or policy rule contingent on the EC choice) first, then the EC government chooses its trade and subsidy policies, and finally the EC firm makes its entry decision.

*FTD*, again depending upon *US* trade policy. World surplus increases provided inequality (5) or (7) holds. Examples of both kinds of outcome changes are the following:

*Example:* (Entry in both games 2 and 5): Consider cell (1,2) in Table 4, which corresponds to cell (1,4) in Table 3. Without subsidies, entry worked through protection and led to *PRE* in game 2. With subsidized free trade the EC government can pursue free trade and achieve *FTD* in game 5; it prefers this given its domestic preference.

*Example:* (No entry in game 2, entry in game 5): Suppose sunk costs are so high that unsubsidized entry is never profitable, even with both protection and access. This seems a possibility in the Airbus example. When a subsidy is allowed, the EC can promote subsidized entry. The US always responds with its domestically preferred policy, and the resulting outcome—*FTD* or *USP*—is preferred by the EC government if duopoly consumer surplus is large enough relative monopoly consumer surplus.

### B. Entry Deterrence

Now compare games 5 and 6. From Table 4, we see that subsidies are largely immune to commitments, threats, and promises aimed at deterrence by the US. Deterrence only works when the US, by committing to protection and depriving the EC firm of duopoly profits in the US market, makes subsidized entry unattractive for the EC. Since such deterrence shifts the outcome from *FTD* to *MON*, it is undesirable provided inequality (5) holds.

*Example:* (Entry in game 5, deterrence in game 6): Return to cell (1,2) in Table 4. The EC chooses subsidized free trade in game 5, and the outcome is *FTD*. In game 6, if the US chooses protection (or in game 7, matching), it leaves the EC government the choice between no subsidized entry (*MON*) and subsidized entry (*USP*). This can go either way, since  $SE_2 > SE_1$  counteracts  $K > PE_2$ . If the choice is *USP*, the US government is better off with *FTD* and so does not make the commitment. But if the choice is *MON*,

it is possible that the *US* government, preferring this to *FTD*, will commit itself.

If the *EC* government's domestically preferred policy is protection, then subsidies do not affect the outcome, except for the following single case:

*Example:* Consider cell (3,2) of Table 3 and see how game 4 works. The *US* can make a commitment to free trade, protection, or matching. The first draws a protectionist *EC* response, and leads to *PRE*. The second removes access and makes protectionist entry-promotion impossible, so the *EC* government follows free trade and accepts *MON*. The third likewise induces the *EC* government to follow free trade and yields *FTD*. The *US* then picks the best of *MON*, *PRE*, and *FTD*. But it clearly prefers *FTD* over *PRE*, leaving a choice between *FTD* and *MON*. Now introduce subsidies, that is, consider game 7. The *EC* government can respond to *US* protection or matching with a policy of subsidy and protection, achieving *AUT*. This is worse for it than *MON*, so the response to protection is unchanged. But it may be better than *FTD*, so the response to matching may change. In that case the *US* no longer has *FTD* available, and since it prefers *MON* to *AUT*, it is left choosing between *MON* and *PRE*. The formal description of the outcome is

$$\text{Max}_{US}(\text{PRE}, \text{MON}, \text{Max}_{EC}(\text{AUT}, \text{FTD})).$$

In effect, we have the possibility that *FTD* is replaced by *PRE*, a change that is harmful to world welfare.

The discussion in this section suggests that, on balance, subsidies are attractive from the *EC*'s viewpoint. They result in greater gains for the *EC* when utilized and they are more difficult for the *US* to deter. From the point of view of world welfare, they are desirable to the extent that they open the *EC* market but undesirable in a few cases where they promote socially suboptimal new entry. On balance, perhaps, they do more good than harm.

## VI. Modifications and Extensions

We conclude the paper by indicating several ways in which the special assumptions of the model can be altered or relaxed. In most cases, the approach we have developed remains valid, but the taxonomy grows even more complex. We do not think it worthwhile to develop an exhaustive catalogue, but in each instance some cases with special claim to realism would be worth pursuing.

1) *Positive Theory of Policy.* We have taken each government's objective to be the sum of its domestic consumers' and producers' surpluses. This can be changed to a weighted sum to reflect distributive concerns; the categories of domestically preferred policies will then depend on the weights.

The case where consumers' surpluses get zero weights deserves special attention, as it reflects the view that policymakers are captives of concentrated producer lobbies. This can be seen as a positive theory of policy as opposed to our normative approach. Formally, it can be handled as a special case of our model. With all consumers' surpluses set to zero, both governments' domestically preferred policy is protection. The first column of Table 3 then shows that *strategic* use of protectionist policies for entry is irrelevant. The only possible exception (game 4 of the bottom cell) also makes no difference: it is not possible for both governments to prefer *FTD* when only profits count, so the outcome is *AUT*.

2) *Unequal Costs.* If the two firms have unequal costs, we cannot specify monopoly or duopoly profits or consumers' surpluses independently of which firm is involved. This increases the number of cases and makes comparisons more complex. A possible use of this generalization will be to shed light on new investments in declining industries. Dynamic games of *R&D* will also involve unequal costs.

3) *More Incumbents.* When there are two or more incumbent *US* firms, protection in the *EC* will replace this oligopoly by an *EC* monopoly, thus reducing consumers'

surplus. This will bias the *EC* domestic preference toward free trade, and reduce the attractiveness of protected entry.

4) *Third Markets*. If we introduce other markets in nonproducing third countries, we add the appropriate monopoly or duopoly profits to the payoffs. This increases the desirability of entry to the *EC*, but for given *K* makes active promotion less necessary. It also increases the desirability of entry deterrence to the *US*, but makes policies towards this aim less effective.

5) *Resales*. We assumed that the two markets were separate, that is, that price discrimination could be maintained without fear of arbitrage through resale by third parties. The structure of international marketing and managerial practices of buyers of products like aircraft make this assumption fairly realistic. It is also common in the literature on oligopolistic trade. However, if we relax it, we have to change the payoffs in the outcomes *MON* and *FTD*, by calculating profits and surpluses in a world monopoly and duopoly with uniform prices. In the other cases, protection effectively separates the markets. Again, the comparisons and classifications become more complicated.

6) *Variable Tariff Rates*. We have allowed only a discrete choice between zero and prohibitive tariffs. If a continuous range of tariff policies is introduced, each country's domestic preference will generate an optimum tariff for the familiar reasons of rent extraction in oligopolistic trade. Strategic considerations of entry will then hinge on: (i) whether profitable entry is consistent with the two countries' domestically optimum tariffs, (ii) whether a higher *EC* tariff can induce entry, (iii) whether a *US* threat of raising its tariff can deter entry promotion, or a promise of lowering its tariff can dissuade it. The principles of the analysis are unchanged, but the details are more complex, and the calculations will have to postulate a particular solution concept for the duopoly.

7) *Access Charges*. The outcomes of the games we have considered do not generally maximize world-social-surplus. First-best world-social-surplus is not maximized be-

cause governments cannot strike bargains with the incumbent monopolist to supply quantities demanded at marginal cost. Second-best world-social-surplus is not maximized because governments cannot strike bargains with the incumbent monopolist or the entrant about access to domestic markets at all. If governments were able to "sell access" to domestic markets to foreign firms for lump sum amounts, presumably much inefficient protectionism would be eliminated. Furthermore, given that foreign firms cannot buy their way into domestic markets directly, they have an incentive to do so indirectly by setting up domestic subsidiaries and transferring enough monopoly rents to these domestic subsidiaries to make a protectionist policy which destroys these rents undesirable. This would tend to overcome some of the nonoptimality of our outcomes.

8) *Repetitions*. We have considered a single occasion where the opportunity for entry arises. If the governments are engaged in such games repeatedly, with the roles of incumbent and entrant falling to different countries on different occasion, then there would be prospects of cooperation by forswearing protection and achieving *FTD* or *MON*, whichever is better.

## REFERENCES

- Brander, James A. and Spencer, Barbara J., "Tariff Protection and Imperfect Competition," in Henryk Kierzkowski, ed., *Monopolistic Competition in International Trade*, Oxford: Oxford University Press, 1984.
- Curtis, Douglas C. A., "Trade Policy to Promote Entry with Scale Economies, Product Variety, and Export Potential," *Canadian Journal of Economics*, February 1983, 16, 109-21.
- Dixit, Avinash, "Recent Developments in Oligopoly Theory," *American Economic Review Proceedings*, May 1982, 72, 12-17.
- , "International Trade Policy for Oligopolistic Industries," *Economic Journal*, March 1984, Suppl., 94, 1-16.
- and Kyle, Albert S., "On the Use of

- Trade Restrictions for Entry Promotion and Deterrence," Economics Discussion Paper No. 56, Woodrow Wilson School, Princeton University, 1983.
- \_\_\_\_\_ and Stiglitz, Joseph E., "Monopolistic Competition and Optimum Product Diversity," *American Economic Review*, June 1977, 67, 297-308.
- Eaton, Jonathan and Grossman, Gene M., "Optimal Trade and Industrial Policy under Oligopoly," Discussion Paper No. 59, Woodrow Wilson School, Princeton University, 1983.
- Grossman, Gene M. and Richardson, J. David, "Strategic Trade Policy: A Survey of Issues and Early Analysis," Princeton University: Special Papers in International Economics, 1984.
- Johnson, Harry G., "Optimum Tariffs and Retaliation," *Review of Economic Studies*, February 1954, 21, 142-53.
- Selten, Reinhard, "A Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Form Games," *International Journal of Game Theory*, January 1975, 4, 25-55.
- Spence, Michael, "Product Selection, Fixed Costs, and Monopolistic Competition," *Review of Economic Studies*, June 1976, 43, 217-35.

# A Framework for Evaluating the Effects of Economic Growth and Transfers on Poverty

By PETER GOTTSCHALK AND SHELDON DANZIGER\*

While the reduction of poverty is a non-controversial policy goal, there is little agreement as to the relative impacts of changes in market and transfer income on poverty. The War on Poverty was predicated on the assumption that increased market incomes resulting from economic growth were not sufficient for alleviating poverty. As the U.S. Council of Economic Advisers stated:

Rising productivity and earnings, improved education, and the structure of social security have permitted many families or their children to escape; but they have left behind many families who have one or more special handicaps. These facts suggest that in the future economic growth alone will provide relatively few escapes from poverty. Policy will have to be more sharply focused on the handicaps that deny the poor fair access to the expanding incomes of a growing economy.

[1964, p. 72]

This view has recently been challenged in public policy debates. According to a recent analysis by the U.S. Office of Management and Budget:

History teaches us that economic growth is a critical determinant of individual and family well-being. In the

decade of the 1970s, the economy failed to perform as well as in the 1960s.... As a result, it was in the 1960s rather than in the 1970s that the greater inroads against poverty were made.

[1983, pp. 30-31]

How sensitive is the incidence of poverty to increased economic activity? One would think that the experience of the last twenty years would offer an almost-ideal social experiment to determine the relative importance of growth in market incomes and income transfers. That economic growth reduces poverty is suggested by the fact that rapid growth during the late 1960's, followed by periods of stagnation and recession, was mirrored by a sharp decline, then a leveling, and then an increase in poverty rates. However, during the period when poverty decreased most quickly, there was also rapid growth in the benefit levels and numbers of recipients of income transfer programs. Furthermore, the recent increase in poverty coincides with reductions in the rate of growth of transfers. In addition, inequality of market income has been increasing since the late 1960's. It is, therefore, impossible to use simple bivariate relationships to account for the sources of changes in poverty.

Most recent attempts (for example, James R. Thornton et al., 1978) to measure the relative importance of growth in market and transfer incomes are in the tradition of W. H. Locke Anderson (1964), Lowell Gallaway (1965, 1967) and Henry Aaron (1967). These studies estimate time-series regressions to obtain the partial effects of economic growth and transfer programs on poverty. Elsewhere, we have shown (1983) that the conflicting conclusions of these studies can be attributed to the time-series data which are too highly collinear to yield robust co-efficient estimates. Minor changes in func-

\*Associate Professor, Department of Economics, Bowdoin College, Brunswick, ME 04011, and Research Associate, Institute for Research on Poverty; and Director, Institute for Research on Poverty, University of Wisconsin, Madison, WI 53706. We thank Daniel Feaster, John Flesher, Michael Kende, Christine Ross, and Nancy Williamson for computational assistance. Robert Haveman, Jonathan Goldstein, Eugene Smolensky, Daniel Weinberg, and the anonymous referees provided valuable comments on an earlier version. Financial assistance from the Alfred P. Sloan Foundation and the University of Wisconsin Graduate School Research Committee are gratefully acknowledged.

tional form or variable measures substantially change the relative importance of market and transfer incomes.

In this paper, we develop an alternative framework for measuring the relative effects on poverty of growth in mean market income, mean transfer income, and inequality. We analyze the relationship between changes in the location and shape of the joint distribution of various income sources (i.e., the mean and higher level moments of market and transfer incomes) and a summary measure of inequality (in this case, poverty). The technique is general and could equally well be applied to other breakdowns of income (for example, husband's and wife's earnings) or other summary measures (for example, quintile shares).

Section I develops the conceptual framework, which is used in Section II to decompose the sources of the decrease in poverty between 1967 and 1979 and the subsequent rise in poverty between 1979 and 1982.

### I. Conceptual Framework

We begin by defining changes in poverty in terms of changes in the underlying distribution of income. Let the poverty rate be defined as

$$(1) \quad P = \int_{-\infty}^T \int_{-\infty}^{T-I_2} g(I_1, I_2, \mathbf{m}) dI_1 dI_2,$$

where  $I_1$  and  $I_2$  are two sources of income (in this case, market and transfer incomes),  $T$  is a poverty threshold, and  $g$  is the joint density function of  $I_1$  and  $I_2$ , which is defined in terms of a vector of parameters,  $\mathbf{m}$ .

Writing the total differential of  $P$  with respect to its  $k$  parameters,

$$(2) \quad dP = \sum_{i=1}^k (\partial P / \partial m_i) dm_i,$$

allows us to decompose the total change in poverty into changes due to each of the  $k$  parameters. Changes in these parameters in turn reflect changes in the moments of the distribution.

To determine the relative effects of observed changes in all the moments of the

distribution on poverty, we must specify a particular functional form for  $g$ . Our distributional assumption is guided by two principles: the function should not be more restrictive than functional forms applied elsewhere in the income distribution literature; and the form should depend on a sufficiently small number of parameters to make empirical work tractable. The displaced lognormal distribution meets these criteria. This three-parameter distribution is more general than the better known two-parameter lognormal distribution. It also corrects for the negative skewness found in the distribution of log income.

Because the official poverty thresholds vary by family size, the bound of integration in equation (1) cannot be treated as a fixed number. To avoid excessive complexity, we divide each household's income by its poverty threshold and define the income-to-needs ratio as

$$(3) \quad I^* = (I_1 + I_2)/T.$$

Any household with an income-to-needs ratio below one is poor.<sup>1</sup>

We assume that  $I^*$  has a displaced lognormal distribution, an assumption supported by a Kolmogorov-Smirnov test (Morris DeGroot, 1975).<sup>2</sup> Therefore,

$$(4) \quad Z = [\ln(I^* + c) - \mu] / \sigma,$$

<sup>1</sup>An alternative solution to dividing income by the poverty line would be to expand the conceptual framework to take account of changes in the trivariate distribution of household needs, market incomes, and transfers. Poverty would then depend on four additional moments (the mean and variance of needs as well as the covariances of needs with market incomes and transfers). While our formulation is less general and requires that we make distributional assumptions about income-to-needs ratios, rather than income alone, it keeps data problems manageable. An examination of the data indicates that the assumption of displaced lognormality was as applicable to income-to-needs as to income.

<sup>2</sup>Since we model changes in poverty in terms of changes in the cumulative distribution below the poverty line, we are concerned with the shape of the distribution at  $I^* = 1.0$ . We performed Kolmogorov-Smirnov tests to see whether we could reject the hypothesis that the income-to-needs distributions were lognormal. Using 1967 and 1982 data on all persons and for persons living

which has a standardized normal distribution with density function  $\phi(Z)$ .<sup>3</sup> The three parameters of this distribution are  $\mu$ ,  $\sigma^2$ , and  $c$ , the displacement factor (John Aitchison and J. A. C. Brown, 1957). Changes in  $c$  reflect the partial effects of changes in the third moment about the mean. The first two parameters are themselves the first two moments of the distribution of  $\ln(I^* + c)$  since

$$(5) \quad \mu = E[\ln(I^* + c)];$$

$$(6) \quad \sigma^2 = \text{var}[\ln(I^* + c)].$$

Since poverty is defined as an income-to-needs ratio less than one, we can write

$$(7) \quad P = \int_{-\infty}^h \phi(z) dz,$$

$$\text{where } h = [\ln(1.0 + c) - \mu] / \sigma.$$

Poverty is, therefore, a function only of the parameters of the distribution of  $\ln(I^* + c)$ .

For ease of interpretation, we use the following results from Charles Metcalf (1972) to define the distribution in terms of  $c$  and the mean ( $\alpha$ ) and variance ( $\beta^2$ ) of  $I^*$ , rather

than the mean and variance of  $\ln(I^* + c)$ :

$$(8) \quad \mu = \ln \left\{ \frac{(\alpha + c)^2}{[\beta^2 + (\alpha + c)^2]^{1/2}} \right\};$$

$$(9) \quad \sigma^2 = \ln \left\{ \frac{[\beta^2 + (\alpha + c)^2]}{(\alpha + c)^2} \right\}.$$

We can, therefore, decompose changes in poverty into three components:

$$(10) \quad dP = (\partial P / \partial \alpha) d\alpha + (\partial P / \partial \beta^2) d\beta^2 + (\partial P / \partial c) dc.$$

This decomposition can be further expanded by recognizing that

$$(11) \quad \alpha = E(I^*) = E(I_1^*) + E(I_2^*);$$

$$(12) \quad \beta^2 = \text{var}(I^*) = \text{var}(I_1^*) + 2\text{cov}(I_1^* I_2^*) + \text{var}(I_2^*),$$

where asterisks again indicate income as a proportion of needs. Therefore,

$$(13) \quad dP = \frac{\partial P}{\partial \alpha} dE(I_1^*) + \frac{\partial P}{\partial \alpha} dE(I_2^*) + \frac{\partial P}{\partial \beta^2} d\text{var}(I_1^*) + 2 \frac{\partial P}{\partial \beta^2} d\text{cov}(I_1^* I_2^*) + \frac{\partial P}{\partial \beta^2} d\text{var}(I_2^*) + \frac{\partial P}{\partial c} dc.$$

Equation (13) shows how the total change in poverty,  $dP$ , can be decomposed into changes in poverty associated with changes in means, the dispersion of each income source, their covariance, and the displacement factor.

Given our distributional assumption, we can derive explicit expressions for the partial derivatives in equation (13):<sup>4</sup>

$$(14) \quad \frac{\partial P}{\partial \alpha} = \frac{\phi(h)}{\sigma^2(\alpha + c)[\beta^2 + (\alpha + c)^2]} \times \{ h\beta^2 - \sigma[2\beta^2 + (\alpha + c)^2] \};$$

<sup>4</sup>The derivation is given in the Appendix.

in households headed by prime-aged men, we tested the hypothesis between the tenth and the fifteenth percentiles for all persons and between the fifth and tenth for men, since these portions of the distribution include all the observed poverty levels for the two groups over the 1967-82 period.

Since the Kolmogorov-Smirnov test is based on the maximum difference between the cumulative empirical and theoretical distributions, large deviations outside the area of interest will lead to the rejection of displaced lognormality for the full distribution. This was in fact the case. However, in the specified ranges, we could not reject at the 5 percent level the hypothesis that the four distributions we tested were lognormal. For example, in 1982, the maximum difference between the two cumulative distributions for prime-age men between the tenth and the fifteenth percentiles was .0025, where the actual value was .0501 and the theoretical value was .0475 percent.

<sup>3</sup>Because we assume that  $I^*$  follows a displaced lognormal distribution, the distribution of market income-to-needs and that of transfer income-to-needs cannot each follow a displaced lognormal—the sum of two displaced lognormals is not itself displaced lognormal. In fact, the empirical distribution of transfers-to-needs is clearly not lognormal.

$$(15) \quad \frac{\partial P}{\partial \beta^2} = \frac{\phi(h)}{2\sigma^2[\beta^2 + (\alpha + c)^2]}[\sigma - h];$$

$$(16) \quad \frac{\partial P}{\partial c} = \frac{\phi(h)}{(1+c)\sigma} + \frac{\partial P}{\partial \alpha}.$$

With time-series data on the displacement factor and the means, variances, and covariance of  $I_1^*$  and  $I_2^*$ , we can use equation (13) and the three partial derivatives in equations (14)–(16) to decompose changes in poverty. The differentials in equation (13) are approximated by finite changes in each parameter. The partial derivatives, which are solely functions of  $\alpha$ ,  $\beta^2$ , and  $c$ , are evaluated at the average values for these parameters for each pair of years.

The methodology developed here differs from that of the time-series regression studies cited earlier. Those studies have looked only at changes in the first moments of the distribution and assumed that the relationship was linear, implying that the derivatives of poverty with respect to the first moments are each constant and that the derivatives with respect to all higher level moments are zero. These unrealistic assumptions, in addition to data problems, cast doubts on the methodology as well as the results of prior studies.

In the following section, we apply our methodology to determine the impacts of growth in market and transfer income on poverty during the period 1967 to 1982.

## II. Empirical Results

The annual March *Current Population Surveys* (CPS) for 1968 through 1983, each of which contains information on about 50,000 households (U.S. Department of Commerce, 1968–83), were used to calculate sample means, variances, and covariance of  $I_1^*$ , which we define as market income (wages, salaries, self-employment income, dividends, interest, rents, private pensions, etc.) measured as a proportion of needs, and  $I_2^*$ , cash government transfers (social insurance and public assistance) measured as a proportion

of needs.<sup>5</sup> This sample information is used to calculate estimates of  $\alpha$  and  $\beta^2$ .

An additional piece of information is needed to estimate  $c$ . Metcalf (p. 21) uses the relationship between  $\alpha$  and the 10 percent and 90 percent cutoffs for that estimation. We follow a similar method. Since we are primarily interested in the shape of the distribution near the poverty cutoff (where  $I^*$  equals one), we use a method which insures that the distribution based on  $\hat{\alpha}$ ,  $\hat{\beta}^2$ , and  $\hat{c}$  yields an estimated cumulative density which is equal to the observed cumulative density at the cutoff.

Let  $h$  be the standardized normal variate, defined in equation (7), which is consistent with the observed poverty rate. Equations (7), (8), and (9) are, therefore, three nonlinear equations in three unknowns— $\hat{\mu}$ ,  $\hat{\sigma}$ , and  $\hat{c}$ —which can be solved from three observed quantities— $\hat{h}$ ,  $\hat{\alpha}$ , and  $\hat{\beta}^2$ . These equations are solved numerically in each year to yield estimates of  $c$ .

The means, variances, covariance, and displacement factor are shown in columns (1)–(6) of Table 1. We use these data to calculate the partial derivatives in each year, shown in columns (7)–(9).<sup>6</sup> The decline in the derivatives of poverty with respect to mean market or mean transfer income in all but recessionary years shows that declining antipoverty effectiveness of growth as the density at the poverty line decreases.

<sup>5</sup>We would prefer to include the value of in-kind market income (for example, fringe benefits) and transfers in our poverty measure. Unfortunately, CPS data tapes that include in-kind transfers are available only for a few years. Since these transfers have been growing, their inclusion would increase the relative antipoverty effects of growth in transfers compared to growth in market income.

We use the terms market income and transfer income interchangeably with the more cumbersome terms market income-to-needs ratio and transfer income-to-needs ratio.

<sup>6</sup>Because changes in the moments of the distribution change the values of the partial derivatives, the derivatives shown in Table 1 are evaluated at the average values of the moments between adjacent years. For example, the 1982 value of  $-0.178$  for the derivative with respect to the means,  $\partial P / \partial \alpha$ , reflects the impact of a unit change in  $\alpha$ , evaluated at the average 1981–82 values of  $\alpha$ ,  $\beta^2$ , and  $c$ .



TABLE 1—MOMENTS, DISPLACEMENT FACTORS, PARTIAL DERIVATIVES, AND POVERTY RATES FOR ALL PERSONS 1967–82

	Mean		Variance		Covariance of Market Income and Transfers	Displace- ment Factor	Partial Derivatives with Respect <sup>a</sup> to			Official Poverty Rate
	Market Income ÷ Needs (1)	Transfer Income ÷ Needs (2)	Market Income ÷ Needs (3)	Transfer Income ÷ Needs (4)			Means (7)	Variances (8)	Displace- ment (9)	
1967	2.546	.155	4.337	.137	-.175	0.254	—	—	—	.143
1968	2.673	.170	4.666	.155	-.195	0.347	-0.216	0.037	0.053	.128
1969	2.794	.173	5.192	.153	-.209	0.372	-0.194	0.032	0.052	.122
1970	2.749	.197	5.201	.179	-.229	0.387	-0.193	0.031	0.054	.126
1971	2.759	.219	5.373	.210	-.261	0.402	-0.191	0.030	0.052	.125
1972	2.923	.239	6.005	.245	-.300	0.620	-0.182	0.029	0.045	.119
1973	2.987	.256	6.112	.283	-.331	0.691	-0.164	0.026	0.039	.111
1974	2.847	.276	5.425	.305	-.341	1.010	-0.180	0.030	0.036	.123
1975	2.756	.314	5.253	.336	-.365	0.899	-0.169	0.030	0.031	.123
1976	2.846	.310	5.525	.354	-.388	0.976	-0.169	0.030	0.030	.118
1977	2.918	.309	5.838	.358	-.399	1.012	-0.161	0.028	0.030	.116
1978	3.007	.307	5.907	.379	-.404	1.363	-0.161	0.029	0.025	.114
1979	2.989	.305	5.762	.376	-.377	1.475	-0.153	0.027	0.023	.117
1980	2.814	.313	5.207	.373	-.349	1.573	-0.166	0.030	0.022	.130
1981	2.792	.309	5.686	.375	-.341	1.159	-0.160	0.027	0.025	.140
1982	2.769	.327	5.950	.404	-.334	1.167	-0.178	0.028	0.029	.150

Source: Computations by authors from March *Current Population Surveys*. Zero and negative incomes are included in calculating the moments.

<sup>a</sup>The derivative is applicable to the change in the moments from the previous the indicated year.

The data in Table 1 give the basic information necessary to decompose changes in poverty. For example, from 1981 to 1982 mean market income and mean transfers (each measured as a proportion of needs) changed by  $-0.023$  and  $0.018$ , respectively. Adding these two components of mean income and multiplying the sum by  $-0.178$  ( $\partial P/\partial \alpha$ , shown in col. (7)) shows that changes in the mean of the income/needs distribution (the first two terms in equation (13)) increased poverty by .0009 (i.e., .09 percentage points). This is less than one-tenth of the total one-point rise in the poverty rate over the two years. The impact of changes in the variances, covariance, and displacement factor on poverty are obtained by repeating the calculations for the other terms in equation (13). When all of these terms are summed, they equal the 1.0 point increase in poverty, shown in column (10).

The computations using the data in Table 1 are summarized in Table 2. We aggregate across time periods and sources of change in

poverty. Since 1979 is the last nonrecessionary year, we examine changes in poverty over the subperiods 1967 to 1979 and 1979 to 1982.<sup>7</sup> The first period is marked by economic growth, as market income-to-needs increased by 17.4 percent, or about 1.4 percent per year; the second by cyclical decline, as market income-to-needs declined by 7.4 percent, or about 2.5 percent per year. In the first period, transfer income-to-needs increased by 9.8 percent; in the second, by 7.2 percent.

Since there have been major demographic shifts in the population, the data for all persons may reflect changing household composition as well as changes in the income distribution of each demographic group. Therefore, Table 2 also analyzes changes in poverty for persons living in households

<sup>7</sup>The data in Table 1 are used to calculate the partial effects in each year; the data in Table 2 show these effects summed over each subperiod.

TABLE 2—DECOMPOSITION OF CHANGES IN POVERTY

	Actual Percentage Point Change in Poverty <sup>a</sup> (1)	Estimated Point Change in Poverty Associated with Change in		
		Mean Market Income- to-Needs Ratio <sup>b</sup> (2)	Mean Transfer Income- to-Needs Ratio <sup>b</sup> (3)	Shape of Distribution (4)
All Persons				
1967-79	-2.6	-3.3	-3.5	+4.2
1979-82	+3.3	+0.9	-0.4	+2.8
Young Men <sup>c</sup>				
1967-79	-2.3	-3.4	-0.7	+1.8
1979-82	+5.8	+3.4	-0.6	+3.0
Prime-Aged Men <sup>d</sup>				
1967-79	-2.4	-3.9	-0.8	+2.3
1979-82	+3.1	+0.9	-0.2	+2.4
Elderly Persons <sup>e</sup>				
1967-79	-14.7	-0.1	-21.1	+6.5
1979-82	-0.4	-0.1	-1.7	+1.4

Source: See Table 1.

<sup>a</sup> The sum of the change in cols. (2), (3), and (4) is equal to the actual change shown in col. (1).

<sup>b</sup> Computed on the assumption that higher level moments change in such a way that the coefficients of variation and skewness all remain constant. See fn. 9.

<sup>c</sup> Persons in households with male head less than 25-years old.

<sup>d</sup> Persons in households with male head between the ages of 25 and 64.

<sup>e</sup> Persons in households with male or female head 65 years or older.

headed by young men and prime-aged men, two demographic groups that are expected to benefit most directly from economic growth, and for persons living in households headed by elderly men and women.<sup>8</sup>

Sources of changes in poverty are aggregated into three groups, in columns (2)-(4) of Table 2. Column (2) shows the impact of changes in mean market income, holding the

coefficient of variation and the coefficient of skewness of market income constant. It can, therefore, be interpreted as the effect of changes in the mean, holding relative inequality constant.<sup>9</sup> Column (3) shows the

<sup>8</sup> Unfortunately, we cannot decompose changes in poverty for persons living in households headed by nonaged women or black men and women. Our method of estimating the displacement factor requires the existence of a value consistent with the observed poverty rate and the mean and variance of the group-specific income-to-needs distribution. No such values exist for households headed by nonaged women or blacks, even when they are disaggregated into four age-race cells. The fact that the displaced lognormal distribution fits well at the poverty line for all persons and all men (see fn. 2), but does not provide a good approximation for women or blacks, suggests further research using alternative distributional assumptions.

<sup>9</sup> We begin with initial-year values for the means and the second- and third-level moments of each demographic group (including the covariances and corresponding third-level cross products). We then assume that each household's market and transfer income-to-needs ratios grew at the same rates as the observed increases in the means of the two income sources and calculate values of the second- and third-level moments consistent with these proportional growth rates. This ensures that the coefficient of variation and coefficient of skewness of each income source remain constant. Note that total income inequality will change when each income source grows proportionally unless each grows at the same rate.

Poverty rates consistent with proportional growth in market income are in col. (2) of Table 2. Col. (3) shows the net poverty reduction that results when transfers grow at their observed rate. Col. (4) shows the impact of changes in the shape of the distribution due to the

corresponding impact on changes in mean transfer income. The last column shows the impact of the remaining factors which change the shape of the distribution. This grouping reflects our interest in the relative effects on poverty exerted by economic growth, transfer growth, and changes in inequality.<sup>10</sup>

Row 1 of Table 2 shows that the official poverty rate for all persons declined by 2.6 percentage points between 1967 and 1979. Column (2) shows the effect of increases in the mean of market income (holding the shape of the distribution constant). This reflects the poverty-reducing effects of increases in mean market income and the poverty-increasing effects of increases in the variance, and the effects of changes in the covariance and displacement factor, all resulting from proportional growth in the market income of each household. The net effect was a decline in poverty of 3.3 percentage points. Increases in mean transfers decreased poverty by an additional 3.5 points. Growth in mean transfer income was, therefore, slightly more important than that of mean market income in reducing poverty over this period.

These two poverty-reducing factors were partially offset by changes in the shape of the distribution, which increased poverty by 4.2 points. Increased inequality was more important than growth in the mean of either income source. Thus, a substantial portion of the antipoverty impact of economic growth was canceled by rising inequality.

---

deviations of the variance, covariance and displacement factor from the values consistent with proportional growth in each household's market and transfer income. The data in col. (4) are computed as residuals, i.e., as the actual values in col. (1) less the computed values in cols. (2) and (3).

<sup>10</sup>The decomposition of Table 2 proceeds as if transfer growth has no effect on the distribution of market income, and vice versa. However, because of labor supply responses to transfers, some portion of the effect of growth in transfers on poverty is being classified as a change due to the resulting reduction in market incomes. Similarly, since some transfers are income-tested, a portion of the effect of growth in market income on poverty is classified as a change due to the resulting decline in transfer income. A simulation that assumed large behavioral responses (see our earlier paper) did not alter our qualitative conclusions.

Between 1979 and 1982, poverty for all persons increased from 11.7 to 15.0 percent. (The decline in mean market income increased poverty by 0.9 points.) The poverty-reducing impact of growth in mean transfers —0.4 points—was much smaller in this period, because transfers as a percentage of needs grew much more slowly after 1979 than in the previous period. Changes in the shape of the distribution increased poverty by 2.8 points in only three years. This confirms the importance of focusing on the shape as well as the position of the distribution—if each household's market income had decreased by the observed 7.4 percent and the transfer income of each had increased by the observed 7.2 percent, poverty would have risen by only 0.5 points (col. (2) plus (3)), instead of the actual 3.3 point increase.

As expected, the relative importance of changes in mean market incomes is greater for persons living in households headed by young and prime-aged men.<sup>11</sup> The impacts of changes in the shapes of the demographic-specific distributions, shown in column (4), were, however, nearly as important for each of these groups as for all persons. The tendency toward greater inequality is, therefore, not solely a reflection of demographic shifts (also, see Martin Dooley and Gottschalk, 1984). Note that the 1979–82 downturn in mean market income and the increased inequality more than offset the observed declines in poverty between 1967 and 1979 for these groups.

The largest drop in poverty between 1967 and 1979 (14.7 points) and the only decline between 1979 and 1982 occurred for households headed by elderly persons (men and women). As might be expected, this decline was almost solely a result of growth in mean transfers, primarily Social Security benefits.

<sup>11</sup>While we focus on the impact of changes in the *group-specific* moments on each group's poverty rate, our methodology could be expanded to reflect the impact of changes in the moments of the *aggregate* distribution on group-specific poverty rates. One would need to estimate the sensitivity of the group-specific moments to changes in the moments of the overall distribution. These elasticities would depend on the source of the change in the overall distribution.

### III. Conclusions

We have developed a framework that shows how changes in the poverty rate result from changes in the location and shape of the income distribution. By assuming that the income-to-needs ratio follows a displaced lognormal distribution, we are able to quantify the relative importance of changes in mean market incomes, mean transfers, and inequality. During the 1967-79 period, changes in transfers were about as important as increases in market incomes for all persons. Transfers were less important for non-aged men, and very important for elderly. Since 1979, the decline in market incomes has increased poverty, and the antipoverty effect of transfer growth has declined. Rising inequality increased poverty for all persons and for each group analyzed during the period of growth as well as during the recent recessionary period.

### APPENDIX

The derivatives of poverty with respect to  $\alpha$ ,  $\beta^2$ , and  $c$  can be obtained by recognizing that

$$(A1) \quad \frac{\partial P}{\partial \alpha} = \frac{\partial P}{\partial h} \left[ \frac{\partial h}{\partial \mu} \frac{\partial \mu}{\partial \alpha} + \frac{\partial h}{\partial \sigma^2} \frac{\partial \sigma^2}{\partial \alpha} \right];$$

$$(A2) \quad \frac{\partial P}{\partial \beta^2} = \frac{\partial P}{\partial h} \left[ \frac{\partial h}{\partial \mu} \frac{\partial \mu}{\partial \beta^2} + \frac{\partial h}{\partial \sigma^2} \frac{\partial \sigma^2}{\partial \beta^2} \right].$$

Each component of equations (A1) and (A2) can be obtained from equations (7) to (9) in the main part of the text. Differentiating equation (7) yields

$$(A3) \quad \partial P / \partial h = \phi(h),$$

$$(A4) \quad \partial h / \partial \mu = -1/\sigma,$$

$$(A5) \quad \partial h / \partial \sigma^2 = -h/2\sigma^2.$$

Differentiating equations (8) and (9) yields

$$(A6) \quad \frac{\partial \mu}{\partial \alpha} = \frac{2\beta^2 + (\alpha + c)^2}{(\alpha + c)[\beta^2 + (\alpha + c)^2]},$$

$$(A7) \quad \frac{\partial \sigma^2}{\partial \alpha} = -\frac{2\beta^2}{(\alpha + c)[\beta^2 + (\alpha + c)^2]},$$

$$(A8) \quad \frac{\partial \mu}{\partial \beta^2} = -\frac{1}{2[\beta^2 + (\alpha + c)^2]},$$

$$(A9) \quad \frac{\partial \sigma^2}{\partial \beta^2} = \frac{1}{\beta^2 + (\alpha + c)^2}.$$

Substituting these expressions into equation (A1) and (A2) yields

$$(A10) \quad \frac{\partial P}{\partial \alpha} = \frac{\phi(h)}{\sigma^2(\alpha + c)[\beta^2 + (\alpha + c)^2]} \times \{h\beta^2 - \sigma[2\beta^2 + (\alpha + c)^2]\},$$

$$(A11) \quad \frac{\partial P}{\partial \beta^2} = \frac{\phi(h)}{2\sigma^2[\beta^2 + (\alpha + c)^2]}[\sigma - h].$$

This shows how poverty changes with the mean and variance of the distribution of  $I^*$ .

The impact of changes in the displacement factor,  $c$ , on poverty can be obtained by recognizing that

$$(A12) \quad \partial h / \partial c = 1/(1 + c)\sigma + \partial h / \partial \alpha.$$

Therefore,

$$(A13) \quad \partial P / \partial c = \phi(h)/(1 + c)\sigma + \partial P / \partial \alpha.$$

### REFERENCES

- Aaron, Henry, "The Foundations of the 'War on Poverty' Reexamined," *American Economic Review*, December 1967, 57, 1229-40.
- Aitchison, John and Brown, J. A. C., *The Lognormal Distribution*, New York: Cambridge University Press, 1957.
- Anderson, W. H. Locke, "Trickle-Down: The Relationship between Economic Growth and the Extent of Poverty among American Families," *Quarterly Journal of Economics*, November 1964, 78, 551-24.

- DeGroot, Morris H., *Probability and Statistics*, Reading: Addison-Wesley, 1975.
- Dooley, Martin and Gottschalk, Peter, "Earnings Inequality Among Males in the U.S.: Trends and the Effect of Labor Force Growth," *Journal of Political Economy*, February 1984, 92, 59-89.
- Gallaway, Lowell E., "The Foundations of the 'War on Poverty'," *American Economic Review*, March 1965, 55, 123-31.
- \_\_\_\_\_, "The Foundations of the 'War on Poverty': Reply," *American Economic Review*, December 1967, 57, 1241-43.
- Gottschalk, Peter and Danziger, Sheldon, "Changes in Poverty, 1967-1982: Methodological Issues and Evidence," Discussion Paper No. 737-83, Institute for Research on Poverty, 1983.
- Metcalfe, Charles, *An Econometric Model of Income Distribution*, Chicago: Rand-McNally, 1972.
- Thornton, James R., Agnello, Richard J. and Link, Charles R., "Poverty and Economic Growth: Trickle Down Peters Out," *Economic Inquiry*, July 1978, 16, 385-93.
- U.S. Council of Economic Advisers, *Economic Report of the President, 1964*, Washington: USGPO, 1964.
- U.S. Department of Commerce, Bureau of the Census, *Annual March Current Population Survey Computer Tapes*, 1968-83.
- U.S. Office of Management and Budget, "Means-Tested Individual Benefits," in *Major Themes and Additional Budget Details: FY 1984*, Washington: USGPO, 1983.

# A Theory of Two-Tier Labor Markets in Agrarian Economies

By MUKESH ESWARAN AND ASHOK KOTWAL\*

Economic analysis of agricultural tenancy has yielded rich insights into the institutional mechanisms that evolve as rational responses to the state of market development and production technology. In many respects, the study of tenancy has been a forerunner of the modern literature that is attempting to create a theory of organization based on the analysis of incentive mechanisms underlying the contractual structure. It may be quite fruitful, therefore, to study premodern institutions, especially if they have recurred in diverse environments or at different time periods, and have proved to be historically tenacious. The more anomalous they seem, at first glance, the more rewarding may their analysis prove to us.

One such institution that has not been subjected to economic analysis until recently (Alan Richards, 1979; Pranab Bardhan, 1983) is the institution of permanent workers. Permanent workers (alternatively referred to as tied laborers, estate laborers, farm servants, or attached workers) have existed in agrarian economies as diverse as those of thirteenth-century England, Tokugawa Japan, East Elbian Germany (1750–1860), the Egyptian Delta (1850–1940), pre-1930 Central Chile, and present day India. This institution has exhibited certain common features across different time periods and regions. First, in sharp contrast to the so-called “casual workers” hired on a daily basis, permanent workers are engaged on long-term contracts that span entire crop periods, years, and, sometimes, lifetimes. Second, the employment relationship between the landlords and

these laborers is highly personalized and involves patronage benefits such as homesteads, consumption credit, holiday gifts, and emergency aid in return for total loyalty. A permanent worker is expected to remain loyal to the landlord and further the landlord's interests even in periods of strife between the landlord and casual workers (Sheila Bhalla, 1976; Bardhan and Ashok Rudra, 1981; Richards, 1979). Third, the incidence of this seemingly backward institution appears to increase in response to what may be construed as modernizing stimuli. The opening up of new markets for Chilean agrarian products in the nineteenth century and the consequent increase in labor demand resulted in an increase in the number and proportion of permanent labor contracts (Richards, 1979). Those regions in North India (Haryana) with wider diffusion of new technology and consequently higher labor demand also exhibit greater proportion of permanent labor contracts (Bhalla). A theory of the institution of permanent labor should, therefore, simultaneously explain: (a) why the landlord places such a premium on loyalty, (b) the choice of the instruments he uses to elicit such loyalty, and (c) the increase in the incidences of permanent labor contracts in response to an increase in labor demand.

Bardhan (1983) has recently proposed an explanation for the institution of permanent labor, based on the following idea. Risk-averse workers faced with an uncertain spot wage can engage in long-term contracts with risk-neutral landlords for a prenegotiated wage, albeit at a rate lower than the expected spot rate. Workers, who are assumed to have heterogeneous opportunity incomes, self-select into the permanent and casual labor markets. The main comparative static result of this model explains the well-acknowledged empirical finding that the proportion of permanent workers is higher in tighter la-

\*Department of Economics, University of British Columbia, Vancouver, B.C., V6T 1Y2 Canada. We thank R. Allen, C. Archibald, P. Bardhan, B. C. Eaton, J. Kesselman, T. Lewis, G. MacDonald, M. Manove, P. Neher, P. Tandon, and anonymous referees for helpful comments.

bor markets.<sup>1</sup> In an earlier paper (1979a), Bardhan proposed an alternative explanation for the existence of permanent contracts that was based on differential recruitment costs. Although he noted the importance of the patron-client aspects of the institution of permanent labor, the focus of his two models was to explain the longer duration of the contract. Patron-client aspects, such as loyalty, which are distinctive and inalienable features of the institution of permanent labor, are yet to be formally analyzed.

In this paper we follow the lead of Richards (1979), who has analyzed the institution of permanent labor in the widely different agrarian economies of East Elbian Germany, Egypt, and Chile. His investigation led him to the hypothesis that this institution emerged as a subtle means of supervising labor. A cursory examination of the differences in the tasks assigned to the two kinds of hired labor reveals that important tasks that require judgement, discretion, and care (and are difficult to monitor) are seldom, if ever, assigned to casual workers.<sup>2</sup> Permanent workers, on the other hand, are often entrusted with such responsibilities, almost as if they were family members. Our theory of the institution of permanent labor is based on the hypothesis that it is an attempt by the landlords to transform hired labor into workers whose behavior would approximate that of family labor, thus reducing the burden of on-the-job supervision. Do any of the stylized facts available on the terms of permanent contracts suggest a mechanism that could elicit such behavior from hired workers?

A significant and yet puzzling observation reported by Prafulla Sanghavi (1969) and Bardhan (1979a) is that permanent workers in Indian agriculture typically enjoy a significantly higher annual income (despite a lower daily wage) than casual workers.<sup>3</sup> In addition, permanent workers get consumption loans, homesteads, and other patronage benefits while casual workers face a great deal of uncertainty on the labor markets (Bardhan, 1983; Bhalla). It seems inconceivable that workers close to subsistence and without either employment opportunities or savings could be indifferent between a permanent contract that assures employment and consumption even in slack seasons and a precarious dependence on casual markets. To a worker at subsistence, neither the greater burden of responsibility and more work, nor the distaste for the serf-like existence under the close control of the landlord are reasons compelling enough to render the two types of contracts equivalent in utility.<sup>4</sup> On the other hand, it is equally puzzling that landlords would find it necessary to pay higher-than-opportunity incomes to their permanent workers. It might be natural to presume that the permanent workers are more able and, therefore, earn higher incomes than casual workers. The income differential between the two classes would be explained as ability-rent only if there is no excess supply of able people. In that case, the composition of the labor force would be insensitive to intensification of agriculture, contradicting the observations of Richards (1979).

An explanation of why employers are sometimes found to pay higher-than-oppor-

<sup>1</sup>In addition to Bardhan's own work on East India (1979), this finding has been found to be empirically valid in Chile (pre-1930), East Elbian Germany (1750-1860), and Egypt (1850-1940), as documented in Richards (1979).

<sup>2</sup>See Shigemochi Hirashima (1978, p. 109) for a description of the differential tasks assigned to the two kinds of workers in Pakistan. Also see Thomas Smith (1959) on the tasks performed by permanent workers in Tokugawa Japan and M. M. Postan (1954) for a description of the duties of estate workers in thirteenth-century England.

<sup>3</sup>In Sanghavi (Table 4.7, p. 100), the data on all states in North India, except for Uttar Pradesh, showed a higher annual income for male attached workers by a range of 15-100 percent. Bardhan (1979a) found that the average level of consumption for the family members of permanent workers in Bengal was Rs. 32/month/capita where as it was Rs. 24/month/capita for the family members of casual workers.

<sup>4</sup>For persuasive accounts suggesting that permanent workers are better off than casual workers, see Richards on Egypt (1982, p. 63); Arnold Bauer on Chile (1971, p. 1072).

tunity incomes to their employees has recently been proposed by B. Curtis Eaton and William White (1983). The idea, put simply, is that an income differential maintained over the opportunity income of the worker serves as a monitoring device; any shirking by the worker would invite the threat of getting fired and losing the stream of income differentials.<sup>5</sup> By replacing the income differential with a utility differential and assuming that the opportunity utility of a permanent worker is the expected utility of a casual worker (i.e., assuming an environment with no other employment opportunity), we can adapt the Eaton-White framework to answer the questions we posed earlier. The landlords transform some of the hired laborers into loyal laborers by keeping them at a higher utility level than what they could otherwise attain. The excess demand for permanent jobs thus created is sustained in equilibrium, since it enables the landlords to entrust responsible tasks to an artificially created cadre of loyal workers who would have been prohibitively expensive to supervise otherwise. The wage that minimizes the total labor costs (including wage and supervision costs) is higher than the wage that would minimize the wage costs alone. This framework explains the observation made by Bhalla and Richards (1979) that the permanent workers constitute a class within the class of agricultural workers—the upper tier in an artificially created “two-tiered” labor force. They receive superior benefits and tend to align themselves with their employers under most circumstances. It is important to note, however, that such contracts are viable only if they are long term and if reputation plays an important role so that the fired worker cannot secure another contract soon afterwards.

The above framework is an accurate representation of the institution of permanent labor as described by historians. For example, Arnold Bauer observes that in nine-

teenth-century rural Chile:

Numerically few, the *inquilinos* [permanent workers] were the cream of the rural labour.... This selectivity was made possible by the limited need for estate labour and the lack of alternatives open to the numerous rural families. The good fortune of being accepted on the hacienda was repaid by the *inquilinos* with service and loyalty.

[1975, p. 56]

Our assumption that permanent workers are kept at a higher utility level than casual workers is borne out by the accounts of Richards (1979) on the institution as it prevailed from 1850 to 1940 in Egypt, from 1750 to 1860 in East Elbian Germany, and in pre-1930 Central Chile. Richards also observes: “An Instmann [permanent worker] dismissed for insubordination would quickly find himself among the insecure ranks of *Eigenkatner* and *Einlieger* [casual workers]” (1979, p. 512).

A legitimate question that may be raised at this point is: why doesn't the landlord offer a tenancy contract to the worker? We have explained elsewhere our view that the choice among fixed rental, sharecropping, and fixed wage contracts are influenced by the distribution of certain unmarketed resources across landlords and workers (see our forthcoming article). It is demonstrated there that this, together with the technology and the type of crops, determines the contractual structure that would prevail; even with a linearly homogeneous technology, tenancy contracts will not necessarily obtain. A permanent contract is essentially a wage contract in which the landlord undertakes management and employs a subtle supervision technique that avoids resorting to continuous monitoring, and we model it as such.<sup>6</sup> Such a supervision technique is viable only with long-term contracts since the landlord

<sup>5</sup>A similar idea also appears in Steven Stoft (1980) and, more recently, in Carl Shapiro and Joseph Stiglitz (1984).

<sup>6</sup>Introducing the possibility of tenancy in this model would greatly complicate the formulation, and is a task to be accomplished in future research.



depends on imperfect indicators of the workers' efforts which are gathered almost costlessly as by-products of other management activities. Any meaningful judgment as to whether a worker has been supplying an acceptable level of effort can only be formed after reviewing the accumulated information on the worker's performance over the entire crop season. The landlord is then able to form a judgment on whether or not to fire the worker. For tractability, we shall assume that after the crop has been harvested and counted, the landlord has sufficient information accumulated to know with certainty if the worker has supplied an acceptable level of effort.

To sum up, we postulate that the institution of permanent labor exists in order to facilitate the assignment of important labor tasks to hired labor without having to devote inordinately large amounts of resources to supervision. It enables the landlord to utilize valuable information about the worker's performance that can be costlessly gathered while the landlord is engaged in performing other managerial activities. The permanent worker's income is maintained at a level that renders him a utility sufficiently greater than his opportunity utility that he would choose to supply the acceptable level of effort. Any change in the casual worker's wages, that are determined in a competitive market, results in a corresponding change in the permanent worker's wages.

In Section I, we present a general equilibrium model that incorporates the seasonal nature of agricultural production. The labor market consists of homogeneous workers allocated between permanent and casual workers according to the different tasks assigned. We work out the implications of the model assuming for the workers' utility function a specific form which gives rise to a labor supply function that is consistent with empirical observations. In Section II, we then carry out comparative static exercises and examine the link between the incidence of permanent labor and the different characteristics of the production technology. In the final section, we elaborate on the general

applicability of the essential principle modeled in this paper to discourage morally hazardous behavior.

### I. The Model

We assume that a single crop is produced each year; the crop takes two periods to produce, each period lasting for one-half year. The two periods posited for the production of a crop enable us to capture the variation in the demand for labor and capital over the year. For concreteness, the first period can be viewed as requiring such activities as soil preparation, tilling, sowing, etc., and the second as the period of harvesting, threshing, etc. Typically, the demand for labor and capital is considerably higher in the second period.

We envisage the production process as entailing the use of three inputs: land ( $h$ ), capital ( $K$ ), and labor. It is imperative for our purposes to disaggregate the labor input, and this we do according to the nature of the tasks performed. It is sufficient to consider two broad categories of tasks. Type 1 tasks are those that involve considerable care and judgment (such as water resource management, the application of fertilizers, maintenance of the draft animals and machines, etc.). Such tasks do not lend themselves to easy on-the-job supervision. Type 2 tasks are those that are routine and menial (such as weeding, harvesting, threshing, etc.). Since they involve little discretion, productivity on such tasks can be directly gauged from the extent of the workers' physical activity. In other words, Type 2 tasks are by their very nature easy to monitor. All workers are assumed to have identical abilities. However, even though all workers are drawn from a homogeneous labor force, the tasks to which they are assigned are not necessarily the same.

We draw a distinction between the length of a worker's employment over a period ( $l$ ) and the "intensity" of effort ( $e$ ) with which he applies himself. Efficient performance of a task (either Type 1 or Type 2) requires an effort level  $\bar{e} > 0$ . Since effort is deemed a bad, a worker on a fixed wage will set  $e = 0$

unless he is monitored. We shall take an efficiency unit of labor to be one worker hired for a whole period ( $l=1$ ) at an effort level  $\bar{e}$ . As will be explained below, Type 1 tasks are performed by workers with long-term contracts, while workers hired on the spot market (casual workers) are entrusted with only Type 2 tasks. Empirically, we observe that casual workers are hired mainly in the peak season (i.e., the second period). This is because the tasks to be performed in period 1 are mainly of Type 1 variety—soil preparation, plowing (which entails the use of draft animals or tractors), application of fertilizers, etc. For simplicity we assume that no casual workers are hired in period 1. We let  $L_p$  denote the number of efficiency units of permanent labor employed per period on a typical farm. A permanent worker's contract is over the infinite horizon unless he is found to shirk. We denote by  $L_c$  the number of efficiency units of casual labor employed on the farm in period 2. A casual worker's contract lasts for the whole or part of this period.

We posit that the output,  $q_1$ , of period 1 can be written

$$(1) \quad q_1 = a \min \{ g_1(K_1, L_p), bh \},$$

where  $K_1$  is the amount of capital used in period 1,  $h$  is the amount of land used, and  $a, b > 0$ , and  $g_1(K_1, L_p)$  is a twice continuously differentiable, linearly homogeneous function that is increasing and strictly quasi concave in its arguments. The production function in (1) implies that there is no substitutability between land and the other two factors of production, and that the potential output of the farm is determined entirely by the amount of land.  $g_1(K_1, L_p)$  can be interpreted as an aggregate of the capital and labor inputs in period 1. We assume that labor is an essential input in period 1, that is, that  $g_1(K_1, 0) = 0$  for all  $K_1$ . The parameter  $b$  is introduced to capture land-augmenting technical change, while  $a$  is introduced to simulate Hicks-neutral technical change.

In period 2, the tasks performed by labor are mostly Type 2 variety. We shall assume that in period 2, casual and permanent labor

are perfect substitutes and both will be employed to do Type 2 tasks. Now the output of the second period will depend nontrivially on the activities of the first period. More precisely,  $q_1$  is an intermediate input and we write the second period's output (the final product),  $q_2$ , as

$$(2) \quad q_2 = \min \{ g_2(K_2, L_p + L_c), q_1 \},$$

where  $K_2$  is the amount of capital used in period 2, and  $g_2$  is a twice continuously differentiable, linearly homogeneous function, increasing and strictly quasi concave in its arguments. The motivation for (2) lies in the interpretation of  $q_1$  as the quantity of unharvested crop and  $q_2$  as the quantity of the final product, that is, the harvested and threshed crop;  $q_1$  is thus a natural upper bound on  $q_2$ .

The price of the output is assumed to be exogenously fixed—set in the world market, say—and is normalized to unity. All farmers are assumed to be price takers in the labor and capital markets. For convenience, we assume that all farms are identical. Then in view of the linear homogeneity of (1) and (2), we can aggregate all farmers into a single price-taking farmer. The quantity  $h$  now represents the total arable land in the economy and is assumed fixed;  $L_p, L_c, K_1, K_2, q_1$ , and  $q_2$  can similarly be interpreted as aggregates. The wage rate of a permanent worker is  $w_p$  per period, while that of a casual worker is  $w_c$ . The rental rate on capital equipment per period, assumed exogenous, is  $r_i$ ,  $i=1,2$ . Since the types of capital used in the two periods are not necessarily the same, we can have  $r_1 \neq r_2$ .

#### A. Demand Side

We now turn to the optimal choices of  $L_p$ ,  $L_c$ , and  $K_i$ ,  $q_i$ ,  $i=1,2$ . Consider the production of a typical crop. First note that the optimal choice of factor inputs in period 2 depends on  $L_p$  and the decisions of the first period. The landlord's decision making must thus be foresighted and must be made with full awareness of how the choice of  $L_p$  and his period 1 decisions will impinge on period

2's choices. In what follows we shall adopt the convention that all expenses (wages and rentals) are incurred at the end of the period.

Given the nature of the production functions and the assumption of a constant and exogenously determined price for the final product, it follows that if production is at all viable, as we assume it is, it is profitable to cultivate all of the arable land. The profit-maximizing output levels in the two periods are

$$(3) \quad q_1 = q_2 = abh.$$

Without loss of generality we shall set  $h=1$ . The factor inputs will thus be determined so as to minimize the total present value cost of producing the outputs in (3). Since the landlord's choices of capital and casual labor are dependent on the amount of permanent labor hired, we first determine his demands of  $K_1$ ,  $K_2$ , and  $L_c$  conditional on his choice of  $L_p$ .

Define the cost functions

$$(4) \quad C_2(q_2, r_2, w_c) \\ \equiv \min_{K_2, L_a} \{ r_2 K_2 + w_c L_a \mid g_2(K_2, L_a) \geq q_2 \},$$

where  $L_a \equiv L_p + L_c$  is the aggregate amount of labor used in period 2, and

$$(5) \quad C_1(L_p, q_1/a, r_1) \\ \equiv \min_{K_1} \{ r_1 K_1 \mid g_1(K_1, L_p) \geq q_1/a \}.$$

At the profit-maximizing output levels given by (3), Shephard's Lemma yields the following factor demands:

$$(6a) \quad K_1^d(L_p, b, r_1) = \frac{\partial C_1}{\partial r_1}(L_p, b, r_1),$$

$$(6b) \quad K_2^d(ab, r_2, w_c) = \frac{\partial C_2}{\partial r_2}(ab, r_2, w_c),$$

$$(6c) \quad L_a^d(ab, r_2, w_c) = \frac{\partial C_2}{\partial w_c}(ab, r_2, w_c).$$

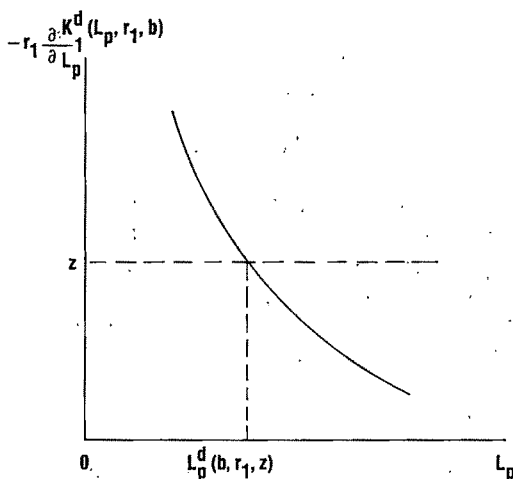


FIGURE 1. DETERMINATION OF THE DEMAND FOR PERMANENT LABOR

The casual labor demand is thus given by

$$(6d) \quad L_c^d(ab, L_p, r_2, w_c) \\ = \max \{ L_a^d(ab, r_2, w_c) - L_p, 0 \}.$$

The optimal choice of  $L_p$  is now determined as the solution to

$$(7) \quad \min_{L_p} r_1 K_1^d(L_p, b, r_1) + \beta r_2 K_2^d(ab, r_2, w_c) \\ + (1 + \beta) w_p L_p + \beta w_c [L_a^d(ab, r_2, w_c) - L_p],$$

assuming that the amount of casual labor hired is strictly positive. In (7)  $\beta$  ( $0 < \beta < 1$ ) denotes the per period discount factor. The first-order condition associated with (7) is

$$(8) \quad -r_1 \left( \frac{\partial K_1^d}{\partial L_p} \right) (L_p, b, r_1) \\ = (1 + \beta) w_p - \beta w_c \equiv z.$$

The demand for permanent labor,  $L_p^d(b, r_1, z)$ , is implicitly determined as the solution to (8). Twice continuous differentiability and the strict quasi concavity of  $g_1(K_1, L_1)$  implies that the left-hand side of (8) is declining in  $L_p$ . Thus  $L_p^d$  is decreasing in  $z$  (see Figure 1).

Together,  $L_p^d(b, r, z)$  and the expressions (6a)–(6d) constitute the demand side of our model. We now turn to the supply side.

### B. Supply Side

Consistent with Bardhan's (1979b) empirical evidence that the agricultural labor supply exhibits low elasticity in the peak period, though it may be fairly elastic in the slack period, we posit the utility function of an agricultural worker to be of the form

$$(9) \quad U(y, e, l) = (y - el)^\gamma; \quad 0 < \gamma < 1,$$

where  $y$  is the income received for the period and  $l$  is the fraction of the period for which he is employed ( $l=1$  if he is hired for the entire period). For an arbitrarily given  $e$  and wage rate  $w$ , the supply response  $l^*(w, e)$ , of a worker is obtained as the solution to

$$(10) \quad \max_l U(wl, e, l) \quad \text{such that } l \leq 1.$$

The maximization in (10) yields the labor supply response:

$$(11) \quad l^*(w, e) \begin{cases} = 0 & \text{for } w < e \\ \in (0, 1) & \text{for } w = e \\ = 1 & \text{for } w > e, \end{cases}$$

and an indirect utility function

$$(12) \quad V(w, e) = \{(w - e)l^*(w, e)\}^\gamma.$$

Since  $V$  is a decreasing function of  $e$ , there is an obvious moral hazard problem under a fixed wage contract, which makes the monitoring of effort imperative. Since Type 2 tasks are easy to monitor, we shall assume that workers performing these tasks can be costlessly supervised. There is thus little reason to hire these workers on long-term contracts, and the conventional means of hiring them, namely, on the spot markets, serves adequately.

With workers performing Type 1 tasks the situation is, however, quite different. We have defined Type 1 tasks as those that involve some discretion and judgment, and are dif-

ficult to monitor. Our discussion in the introduction leads to the following view on the nature of contracts given to workers performing Type 1 tasks.

In order to provide a self-enforcing (incentive) contract, the landlord offers Type 1 workers a permanent contract (over the infinite horizon),<sup>7</sup> in which the worker receives a wage  $w_p$  per period in exchange for the worker's services for the fraction  $l^*(w_p, \bar{e})$  of each period at an effort level  $\bar{e}$ . The worker's effort in period 1 is assumed to be accurately imputable at the end of the year. If the worker is found to have shirked, he is fired at the end of the crop.<sup>8</sup> He is, however, paid his wage,  $w_p$ , for each of the two periods. Once a Type 1 worker is fired, he cannot be rehired except as a casual worker.<sup>9</sup> If  $w_p$  is high enough that a worker's increase in utility from shirking in one period is more than offset by the discounted loss in his utility in having to join the casual labor force, he would never shirk.

It is important to spell out the terms required for the viability of such a contract to permanent workers. First, since a permanent worker's effort can be gauged only at the end of the second period, he can be fired only at the end of period 2. If the landlord concludes that the worker has shirked and decides to fire him, he must still be contractually committed to pay him the prenegotiated wage  $w_p$  in each of the two periods. Without such a contractual commitment, the landlord cannot be trusted to pay even honest workers

<sup>7</sup>While the assumption of infinite time horizon is analytically convenient, it is also empirically appropriate when a permanent worker's status can be inherited.

<sup>8</sup>It might be argued that the landlord would be indifferent between retaining the disloyal worker and replacing him with another who has exactly the same propensity to shirk. A permanent worker who realizes this cannot be deterred from shirking. Since the credibility of the system is at stake, however, the landlord would strictly prefer to replace the disloyal worker, establishing his reputation as a firm enforcer of contracts.

<sup>9</sup>This is assumed for simplicity. For our purposes it is enough if he can secure another such contract only with a probability that is strictly less than unity. This would lead to a discretely lower present value expected earning if he is fired.

and no worker would accept the contract. It may be argued that the fear of notoriety and the consequent difficulty in finding labor would keep the landlord honest. We feel, however, that in a labor surplus economy, reputation can hardly be as effective a check on the behavior of the owner of the scarce factor (land) as it is on the behavior of laborers competing for permanent jobs. Reputation is an effective weapon against moral hazard only for the suppliers of those factors that are in excess supply. The method of eliciting the desired level of effort from an employee by keeping him over his opportunity utility serves precisely to create such an excess supply so that reputation matters.

Another valid question is why a casual worker who failed to secure a permanent contract does not entice the landlord into offering him such a contract by posting a bond, the present value of which is marginally less than the difference in the present values of the lifetime income streams of the permanent and casual workers. Once again, such an arrangement is not viable due to the possibility of moral hazard on the part of the landlord; he always has the incentive to claim at the end of the period that the worker has shirked and thus expropriate the bond. Besides, as Eaton and White (1982) have pointed out, a worker faced with asset constraints may be unable to raise the amount necessary to post such a bond.

We can determine  $w_p$  in terms of  $w_c$  as follows. Assuming, for simplicity, that workers discount their utility at the same rate  $\beta$  as the landlord discounts profits, the present value utility of a permanent worker who is honest (i.e., who never shirks) is given by<sup>10</sup>

$$(13) \quad J_p^h(w_p, \beta) = V(w_p, \bar{e}) / (1 - \beta).$$

Now the opportunity utility of a permanent worker is the utility he would receive as a casual worker. Assume that the casual labor demand is spread uniformly across all the casual workers. Then the discounted lifetime

utility of a casual worker is given by

$$(14) \quad J_c(w_c, \beta) = (\beta / (1 - \beta^2)) V(w_c, \bar{e}).$$

We now turn to the possibility of shirking on the part of a permanent worker. Since any shirking is guaranteed to result in termination at the end of the second period of the same crop, a permanent worker who chooses to shirk will find it optimal to set  $e = 0$  in the first period. Since in period 2 he performs only menial tasks, which can be costlessly monitored, shirking is not possible. His discounted utility over this crop (relative to the beginning of the crop) is

$$V(w_p, 0) + \beta V(w_p, \bar{e}).$$

Further, assuming demand and supply conditions to be identical across all years, a permanent worker who contemplates shirking will do so in the very first year. Thus the discounted lifetime utility of a permanent worker who shirks is

$$(15) \quad J_p^s(w_p, w_c, \beta) = V(w_p, 0) + \beta V(w_p, \bar{e}) + \beta^2 J_c(w_c, \beta).$$

To ensure that a permanent worker never shirks, we simply require

$$(16) \quad J_p^h(w_p, \beta) \geq J_p^s(w_p, w_c, \beta).$$

For given  $w_c$  and  $\beta$ , inequality (16) puts a lower bound on the permanent worker's wage,  $w_p$ , which will elicit the required level of effort. At any  $w_p$  that satisfies (16) a worker obtains a strictly higher utility in a permanent contract than in a series of spot contracts:<sup>11</sup>

$$(17) \quad J_p^h(w_p, \beta) > J_c(w_c, \beta).$$

<sup>11</sup> This can be seen by rewriting inequality (16) as

$$(1 - \beta + \beta^2) V(w_p, \bar{e}) \geq (1 - \beta) V(w_p, 0) + \frac{\beta^3}{1 + \beta} V(w_c, \bar{e}) > (1 - \beta) V(w_p, \bar{e}) + \frac{\beta^3}{1 + \beta} V(w_c, \bar{e}),$$

by (12), so that  $(1 + \beta) V(w_p, \bar{e}) > \beta V(w_c, \bar{e})$ .

<sup>10</sup> This assumes that there is no saving, so that consumption and income are identical.

It follows that the number of permanent workers hired will be demand determined in general. Since a laborer strictly prefers being a permanent worker to being a casual worker, there will generally be an excess supply of workers seeking permanent contracts. This, however, will not result in a downward pressure on the permanent workers' wage, since any wage which is lower than the smallest  $w_p$ , say  $\bar{w}_p(w_c, \beta)$ , that satisfies (16) for given  $w_c$  and  $\beta$  is not credible: it leaves an incentive for the permanent worker to shirk. A casual worker who seeks to obtain a permanent contract by offering to work for a wage marginally less than  $\bar{w}_p$  will find that the landlord will not entertain the offer.

In the next section, we shall find that the behavior of  $\bar{w}_p(w_c, \beta)$  as a function of  $w_c$  is of crucial importance in determining the response of the agricultural economy to various exogenous changes. This behavior is recorded in the following proposition.

**PROPOSITION 1:** *For  $w_c \geq \bar{e}$ , an increase in  $w_c$  warrants a change in  $w_p$ , that is, (a) positive, and (b) if  $\bar{w}_p(w_c, \beta) < w_c$ , then*

$$(18) \quad d\bar{w}_p/dw_c < \beta/(1 + \beta).$$

(Proof: See the Appendix.)

Part (a) of Proposition 1 is eminently reasonable, since an increase in  $w_c$  amounts to an increase in the permanent worker's opportunity income (and utility). According to part (b), when the permanent worker's per period wage rate  $\bar{w}_p(w_c, \beta)$  is less than that of a casual worker's,  $w_c$ , the increase ( $\Delta w_p$ ) that is required to compensate a permanent worker for an exogenous increase ( $\Delta w_c$ ) in a casual worker's wage rate satisfies the inequality

$$(19) \quad (1 + \beta)\Delta w_p - \beta\Delta w_c < 0.$$

This implies that the increase in present value cost of engaging a permanent worker is less than that of a casual worker.

We now turn to the determination of the equilibrium. The equilibrium levels of capital in the two periods are demand determined.

Since permanent workers are held above their opportunity utilities, their number,  $L_p^*$ , is also demand determined:

$$(20a) \quad L_p^*(b, r_1, z) = L_p^d(b, r_1, z).$$

The demand for casual workers, we have seen, is given by

$$(20b) \quad L_c^d(L_p, ab, w_c, r_2) = L_c^d(ab, r_2, w_c) - L_p^*(b, r_1, z),$$

assuming the demand to be strictly positive. Next, we have the condition (16), which translates into

$$(20c) \quad V(w_p, \bar{e})/(1 - \beta) \geq V(w_p, 0) + \beta V(w_p, \bar{e}) + (\beta/(1 - \beta^2))V(w_c, \bar{e}).$$

For any  $w_c$ , (20c) determines the minimum  $w_p$  that will prevent a permanent worker from shirking.

Note that an equilibrium must have  $w_c \geq \bar{e}$  and  $w_p \geq \bar{e}$ , in view of (11). Note also that  $w_p = \bar{e}$  is never a solution to (20c) when  $w_c \geq \bar{e}$ . Thus we must have  $w_p > \bar{e}$ , and consequently,  $L^*(w_p, \bar{e}) = 1$  for a permanent worker. In other words, each permanent worker provides one efficiency unit of labor per period. Let  $N$  be the (exogenously given) total number of workers in the agrarian economy. The aggregate supply of casual labor in the second period,  $L_c$ , is then given by

$$(20d) \quad L_c^s \begin{cases} = 0, & \text{for } w_c < \bar{e} \\ \in (0, N - L_p^*) & \text{for } w_c = \bar{e} \\ = N - L_p^* & \text{for } w_c > \bar{e}. \end{cases}$$

This completes the specification of our model. Exogenous to the model are the production and utility functions, the discount factor, the rental rates on capital, and the total labor force. Endogenous to the model are the wage rate of the permanent and casual workers, the number of permanent workers, the number of efficiency units of

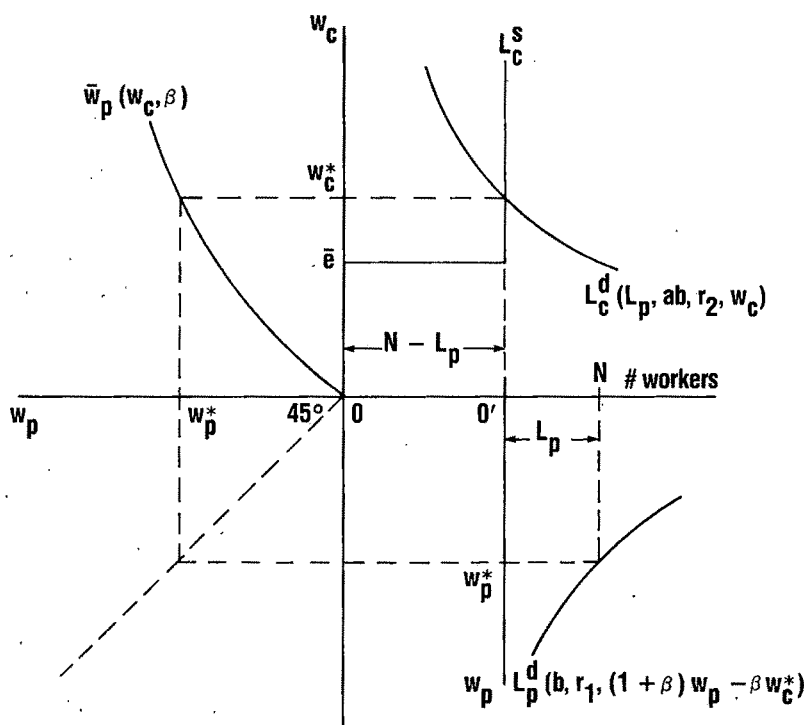


FIGURE 2. AN EQUILIBRIUM WITH UNEMPLOYMENT IN PERIOD 1 AND FULL EMPLOYMENT IN PERIOD 2

casual workers hired in the second period, and the amounts of capital hired in each of the two periods. These are obtained as the solution to the general equilibrium system defined by (20a)–(20d). The employment of capital is demand determined, that is, by (6a) and (6b).

Since a permanent worker's contract extends over the infinite horizon, the hiring of a permanent laborer represents a sunk cost for the landlord. The choice of the labor mix between permanent and casual workers can thus be viewed as a choice between sunk and variable costs.

For an arbitrarily chosen value of  $L_p$ , the casual labor supply is given by the kinked curve  $L_c^s$  in Figure 2. The demand for casual labor, contingent on the choice of  $L_p$ , is obtained from (20b) and is also shown in the first quadrant of Figure 2. The casual labor market clears at the wage rate  $w_c^*$ . (In what follows stars denote equilibrium values.) The second quadrant displays the solution for  $w_p$

in terms of  $w_c$  as obtained from (20c). Associated with a casual labor wage rate  $w_c^*$  is a permanent labor wage rate  $w_p^*$ . The fourth quadrant displays the demand for permanent labor as a function of  $w_p$  when the casual labor wage rate is  $w_c^*$ . For convenience this demand for permanent labor is measured from  $O'$  (along the horizontal axis). If we have indeed located an equilibrium, the demand for permanent labor at  $w_p^*$  will be exactly equal to the  $L_p$  with which we began our construction. Thus the situation illustrated in Figure 2 represents an equilibrium of the system of equations (20a) through (20d).

Given our assumption that the number of casual workers hired is strictly positive, two distinct situations can emerge as equilibria, although both of these are not equally relevant:

$$\text{Case 1: } 0 < L_p^* < N; 0 < L_c^* < N; \\ L_p^* + L_c^* < N.$$

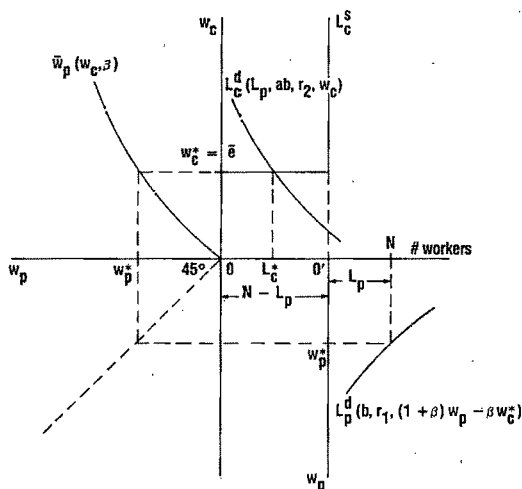


FIGURE 3. AN EQUILIBRIUM INVOLVING UNEMPLOYMENT IN PERIODS 1 AND 2

This situation is illustrated in Figure 3. The demand for casual labor in the peak season is not enough to warrant full employment: there is some unemployment at the equilibrium casual wage rate  $w_c^* = \bar{e}$ .

Case 2:  $0 < L_p^* < N$ ;  $0 < L_c^* < N$ ;  
 $L_p^* + L_c^* = N$ .

This is the situation we considered in Figure 2. Here the supply of labor is a binding constraint in the second period and  $w^* > \bar{e}$ .

Of these two cases, the empirically relevant one is Case 2—in which only part of the labor force is hired on a permanent basis, while in the peak season there is no unemployment. In what follows, therefore, we shall focus exclusively on this case.

## II. Results

We now turn to the comparative static results of our model. These results depend crucially on whether the casual wage rate exceeds or falls short of the wage rate of the permanent workers. These are, of course, endogenously determined and our model allows for both possibilities. However, since our purpose here is to confront our predictions with what empirical evidence there is,

we pursue the empirically relevant case. From Richards (1979), Sanghavi, Ashok Rudra (1982), and Rakesh Basant (1984), we gather this case to be one where

$$(21) \quad w_c^* > w_p^*.$$

In what follows we shall assume (21) to be true.<sup>12</sup> (The signs of the comparative static results below are reversed when (21) is violated.) Defining

$$z^* = (1 + \beta)w_p^* - \beta w_c^*.$$

we see from (18) that

$$(22) \quad \frac{dz^*}{dw_c^*} = (1 + \beta) \left[ \frac{dw_p^*}{dw_c^*} - \frac{\beta}{1 + \beta} \right] < 0.$$

That is, the difference in the present value cost of hiring a permanent worker over that of hiring a casual worker declines with  $w_c^*$ . This fact is used in establishing the comparative static properties of our model, which are recorded in the following proposition.

**PROPOSITION 2:** *In an equilibrium corresponding to Case 2,*

- (a) *an increase in  $N$  decreases the proportion of permanent contracts,*
- (b) *an increase in  $a$  (or  $b$  or both) increases the number of permanent contracts,*
- (c) *an increase in  $a$ , with  $ab$  held constant, decreases the number of permanent contracts,*
- (d) *an increase in  $r_1$  or  $r_2$  increases the number of permanent contracts.*

(Proof: See the Appendix.)

Parts (a) and (b) of Proposition 2 provide explanations, alternative to Bardhan's (1983), for certain empirical observations on permanent labor. According to (a), the proportion of permanent workers is higher the tighter the labor market. A reduction in the supply of agricultural labor,  $N$ , increases the peak season casual wage rate,  $w_c^*$ . This results in

<sup>12</sup> Full employment in the peak period is a necessary condition for this to hold. It is also necessary that the permanent workers not discount the future too heavily.



an increase in the wage rate of the permanent workers,  $w_p^*$ . However, the increases satisfy inequality (19)—implying that the marginal permanent worker is becoming cheaper to hire relative to a casual worker in period 2—inducing a substitution of permanent for casual workers. Part (a) above explains the dramatic increase in the percentage of permanent contracts in East Prussian agriculture in the first half of the nineteenth century. During this period there was an increase in the cultivated area by almost 90 percent between 1815 and 1849, and a simultaneous agrarian reform resulted in peasants losing land to large landlords. The loss of land forced the peasants into the labor market. Richards (1979), however, estimates that the total net loss of land by the peasants to the landlords may have been as low as 3 percent, implying an overall decrease in the labor-to-land ratio, resulting in a higher proportion of permanent workers.

Part (b) of Proposition 2 implies that a yield-increasing improvement in the technology, either through a Hicks-neutral technical change (i.e., higher  $a$ ) or through a land-augmenting technical change (i.e., higher  $b$ ) would increase the proportion of permanent contracts. The intuition for this is the same as that for part (a), and hinges on the relative changes in the magnitudes of  $w_c^*$  and  $w_p^*$  triggered by the exogenous change. Bardhan (1983) provides empirical evidence based on the second Agricultural Labor Enquiry Data that the percentage of permanent labor in India is positively correlated with the index of land productivity.

The demand for permanent and casual labor is, of course, a derived demand. For simplicity we have assumed that the price of the output is exogenously given. It is clear, however, that any factor that affects the demand for output will have repercussions on the labor composition in equilibrium. In particular, an increase in the output price will induce an increase in the output for production functions more general than the ones we have adopted. The effects of an increase in the price of the output can, however, be simulated in our model by an increase in  $a$ . Part (b) of Proposition 2 then explains the impact of the opening up of export markets

on the labor composition in nineteenth-century Chile. In the 1860's, Chile began to export grain to European markets and this lasted until 1890. Bauer (1971) estimated that the percentage of casual workers in the rural labor force of central Chile fell from 72 percent in 1865 to 39 percent in 1895—an observation that is consistent with the result in part (b) of the above proposition.

While part (b) is of empirical interest since it is easily verifiable, an exercise that is of theoretical interest is contained in part (c). Here the final output is held fixed and the burden of activity is shifted across the two periods. We see that an increase in  $a$ , implying a decrease in  $b$ , makes cultivation less intensive in the first period while increasing the activity in the peak season. Since in the second period casual and permanent labor are substitutable, we observe a shift from permanent to casual labor. Thus Jan Breman (1974) observes that a change in crops from rice (which has a relatively even distribution of tasks over the two periods) to mangoes (which has a very heavy labor demand in period 2) resulted in the replacement of permanent contracts by casual labor contracts in Gujarat, India. Kalpana Bardhan (1977) has also made similar empirical observations.<sup>13</sup>

Part (d) of Proposition 2 indicates that a decrease in the rental cost of the type of capital used in the first period would displace permanent workers and consequently increase the use of casual labor in the second period. It could be argued that in India, in view of the notoriously imperfect capital markets, farms with tractors are those for which the owners face lower capital costs. If tractors were employed on such farms only during period 1 (for operations such as ploughing and sowing), the result would be a displacement of permanent workers and an increase in the use of casual workers. While

<sup>13</sup>Part (c) of Proposition 2 is also consistent with empirical evidence that increases in the cropping intensity, which would result in a more even labor demand profile, are correlated with higher incidence of permanent contracts. See K. Bardhan on India, and Richards (1979) on East Prussia.

the existing empirical literature (Rudra; Bina Agarwal, 1981) bears out our prediction regarding permanent workers, there is conflicting evidence on the effect on the employment of casual workers. We conjecture that this conflict arises because tractors are used on some farms for period 1 operations only, while on others they are also used in period 2.

An interesting feature of the result in part (d) of Proposition 2 is the implied complementarity between the capital used in the two periods. Since there are no sunk costs involved in the use of capital (they are presumed to be rented separately in each period), one might expect the choice of the amount of capital used in period 1 to be independent of  $r_2$ . This, however, is not so. A decline in  $r_2$  increases the demand for  $K_2$  and reduces the demand for casual labor. Given full employment in the second period, this lowers the casual wage rate, which in turn lowers the wage rate of permanent workers. In view of (18), however, permanent workers are becoming relatively more expensive than casual workers and this induces a substitution away from permanent labor. The reduction in the amount of permanent labor hired warrants an increase in  $K_1$  since, in the first period, these two inputs are substitutable. A policy implication of this result is that any governmental effort to alleviate labor-supply bottlenecks in the peak period by lowering  $r_2$  (through subsidies, for example) would have an adverse effect on the employment of labor in the slack period.

### III. Conclusions

In this paper we have presented the view that the institution of permanent workers exists to elicit loyalty and trustworthiness from hired workers, so that they can be entrusted with important tasks that are inherently difficult to monitor. This is accomplished by holding them at a higher-than-opportunity utility, and thus creating in the process two tiers within a homogeneous labor force. Evidence of disloyal behavior (i.e., shirking) results in the termination of the permanent contract and the possibility of

the consequent discrete fall in the utility keeps the worker loyal.<sup>14</sup>

A well-known result in the agency literature states that a threshold contract with a discontinuous reward system can be devised to elicit the optimal amount of effort from a worker. The incentive mechanism implicit in permanent contracts within a two-tiered labor force is, however, not a special case of this. A contract which stipulates the agent's reward in terms of his effort when the verdict on the latter is pronounced *ex post* by the principal will not be accepted by the agent. The possibility of morally hazardous behavior on the part of both the principal and the agent has to be explicitly recognized. In the institution we have discussed, the relationship between the principal and the agent involves repeated transactions and this facilitates the design of a contract which gets around the above difficulty. The landlord is contractually committed to pay the permanent worker the full stipulated income for the year even if he is fired at the end of the year. In other words, the compensation goes with the position; as long as a worker is in the higher tier he receives a compensation appropriate to this position. The contract is thus incentive compatible for both parties despite the inherent problem that there is no objective criterion by which to gauge the worker's effort level.

Long-standing relationships that involve repeated transactions between two parties and put a premium on loyalty and trust are referred to as patron-client relationships.

<sup>14</sup>Note that the assumption of a homogeneous labor force is not essential to our theory. With heterogeneous alternative employment opportunities across workers, casual workers with high opportunity incomes may not prefer permanent contracts. Even so, our theory remains valid as long as there is an excess supply of some casual workers desiring permanent jobs. Bardhan and Rudra (1981) found, in a survey conducted in West Bengal (India), that the bulk of the casual workers preferred casual contracts and the bulk of the permanent workers preferred permanent contracts. However, a statistical test performed on their data leads us to reject the hypothesis that there is no excess supply of workers desiring permanent contracts in favor of the hypothesis of a strictly positive excess supply.

These relationships are often sustained and strengthened by means of implicit contracts. The patron (the principal) maintains the client (the agent) at a higher-than-opportunity utility through patronage to win the client's loyalty and trust. The instruments used to effect this patronage will vary according to the needs of the client and the ability of the patron to supply these needs. Ideally, the instruments will bestow a large benefit on the client at a relatively low cost to the patron. Provision of land plots in labor-scarce economies, consumption credit in an environment of imperfect capital markets or protection in the dubious legal environment of Sicily are examples of instruments of patronage. As discussed in this paper, the seasonal nature of agriculture renders it relatively easy for landlords to maintain a utility-differential between permanent and casual workers.

Even in industrialized economies we observe contracts that resemble those of permanent workers. In particular, in sectors subject to seasonal demand (such as construction, services catering to tourism, recreational vehicle services, etc.) firms retain year-round a core of permanent workers selected from the same pool as the seasonal workers. We further conjecture that the most familiar type of employment contract, namely, the salaried contract of a white-collar worker (in a position that is inherently difficult to monitor), embodies a supervision mechanism similar to the one discussed in this paper.

The institution of permanent workers is a graphic manifestation of the consequences of the supervision principle proposed by Eaton and White (1983). This principle, however, is quite general in scope. For example, it explains a fact that forms the premise of numerous models in the migration literature: the substantial wage differential that exists between newly recruited factory workers and those in the informal urban sector from which they are recruited.

Indeed, this incentive mechanism does not even require that the transaction between the principal and agent be voluntary. The mechanism could work equally well in a slave economy. What is essential is that the agent

be convinced in a credible fashion that there is a state of existence that is discretely worse than his current one. Even the miserable existence of a slave can be made worse by selling him and separating him from his family. More subtle means employed to create a favored status among slaves in the antebellum South are discussed by Robert Fogel and Stanley Engerman (1974). Thus even when crude supervision devices such as physical punishment are permitted by society, subtle incentive mechanisms that reduce the cost of supervision have always played an important role. The study of historical institutions and their incentive structures could, therefore, be quite useful in the construction of a theory of economic organizations.

#### APPENDIX

##### PROOF of Proposition 1:

(a) Substituting (13), (14), and (15) in (16), which holds with equality at  $\bar{w}_p$ , we have

$$(A1) \quad V(\bar{w}_p, \bar{e})/(1-\beta) = V(\bar{w}_p, 0) + \beta V(\bar{w}_p, \bar{e}) + (\beta^3/(1-\beta^2))V(w_c, \bar{e}).$$

Differentiating the above expression totally with respect to  $w_c$  and rearranging, we obtain

$$(A2) \quad \frac{d\bar{w}_p}{dw_c} = \left[ \left[ \beta^3/(1+\beta) \right] \frac{\partial V}{\partial w_c}(w_c, \bar{e}) \right] / \left[ (1-\beta+\beta^2) \frac{\partial V}{\partial w_p}(\bar{w}_p, \bar{e}) - (1-\beta) \frac{\partial V}{\partial w_p}(\bar{w}_p, 0) \right].$$

The numerator of the right-hand side is clearly positive. Note that for (A1) to hold we must have  $\bar{w}_p > \bar{e}$ . Using (11), it follows from (12) that

$$(A3) \quad \frac{\partial V}{\partial w_p}(\bar{w}_p, \bar{e}) > \frac{\partial V}{\partial w_p}(\bar{w}_p, 0),$$

so that from (A2)

$$(A4) \quad d\bar{w}_p/dw_c > 0.$$

(b) From (A3), we see that the denominator of the right-hand side of (A2) exceeds

$$(1 - \beta + \beta^2) \frac{\partial V}{\partial w_p}(\bar{w}_p, \bar{e}) - (1 - \beta) \frac{\partial V}{\partial w_p}(\bar{w}_p, \bar{e}) = \beta^2 \frac{\partial V}{\partial w_p}(\bar{w}_p, \bar{e}).$$

Thus from (A2) we have

$$\frac{d\bar{w}_p}{dw_c} < \frac{\beta}{1 + \beta} \left[ \frac{\partial V}{\partial w_c}(w_c, \bar{e}) / \frac{\partial V}{\partial w_p}(w_p, \bar{e}) \right]$$

If  $\bar{w}_p(w_c, \beta) < w_c$ , it follows by differentiating (12) that the term in the square bracket is less than unity, so that  $d\bar{w}_p/dw_c < \beta/(1 + \beta)$ .

PROOF of Proposition 2:

First, note that  $K_1^d(L_p, b, r_1)$ , which is obtained as the solution to the trivial optimization in (5) with  $q_1 = ab$ , has the following comparative static properties:

$$(A5) \quad \frac{\partial K_1^d}{\partial L_p}(L_p, b, r_1) < 0,$$

$$\frac{\partial K_1^d}{\partial b}(L_p, b, r_1) > 0, \quad \frac{\partial K_1^d}{\partial r_1}(L_p, b, r_1) = 0.$$

The comparative static properties of  $L_p^d(b, r_1, z)$ , obtained by differentiating (8), using (A5) and the strict quasi concavity of  $g_1(K_1, L_p)$ , are easily seen to be given by

$$(A6) \quad \frac{\partial L_p^d}{\partial b}(b, r_1, z) > 0,$$

$$\frac{\partial L_p^d}{\partial r_1}(b, r_1, z) > 0, \quad \frac{\partial L_p^d}{\partial z}(b, r_1, z) < 0.$$

If we let  $\alpha$  denote an exogenous shift param-

eter whose comparative static effects we wish to determine, we may write

$$(A7) \quad \frac{dL_p^*}{d\alpha} = \frac{\partial L_p^d}{\partial \alpha}(b, r_1, z^*) + \frac{\partial L_p^d}{\partial z^*}(b, r_1, z^*) \frac{dz^*}{dw_c^*} \frac{dw_c^*}{d\alpha},$$

recalling that the number of permanent workers is demand determined. From (20b) and (20d), we have

$$(A8) \quad L_a^d(ab, r_2, w_c^*) = N.$$

Totally differentiating this with respect to  $\alpha$  and rearranging, we have

$$(A9) \quad \frac{dw_c^*}{d\alpha} = \left( \frac{dN}{d\alpha} - \frac{\partial L_a^d}{\partial \alpha} \right) / \left( \frac{\partial L_a^d}{\partial w_c} \right).$$

The comparative static properties of  $L_a^d(ab, r_2, w_c)$ , which is obtained as the solution to the optimization problem (4), are easily verified to be

$$(A10) \quad \frac{\partial L_a^d}{\partial r_2}(ab, r_2, w_c) > 0,$$

$$\frac{\partial L_a^d}{\partial w_c}(ab, r_2, w_c) < 0, \quad \frac{\partial L_a^d}{\partial (ab)}(ab, r_2, w_c) > 0.$$

(a) Since  $\partial L_a^d/\partial N = 0$  and  $\partial L_a^d/\partial w_c < 0$ , (A9) yields  $dw_c^*/dN < 0$ . Further, since  $\partial L_p^d/\partial N = 0$ , we have from (A7) and (22) that  $dL_p^*/dN < 0$ .

$$\therefore \frac{d}{dN}(L_p^*/N) = -\frac{1}{N^2}L_p^* + \frac{1}{N} \frac{dL_p^*}{dN} < 0.$$

(b) Since by (A10)  $\partial L_a^d/\partial a > 0$ , we have  $dw_c^*/da > 0$  from (A9), so that from (A7) we have  $dL_p^*/da > 0$ . As above,  $dw_c^*/db > 0$  since  $\partial L_a^d/\partial b > 0$ . Also, since  $\partial L_p^d/\partial b > 0$ , it follows from (A7) that  $dL_p^*/db > 0$ .

(c) When  $a$  changes but  $ab$  is held constant, we see from (A8) that  $dw_c^*/da = 0$ .

Thus from (A7),

$$\text{sign} \left\{ \frac{dL_p^*}{da} \bigg|_{ab = \text{constant}} \right\} = \text{sign} \left\{ \frac{\partial L_p^d}{\partial a} (1/a, r_1, z) \right\} < 0.$$

(d) Since  $\partial L_a^d / \partial r_1 = 0$ , we have from (A9) that  $dw_c^* / dr_1 = 0$ .

$$\therefore dL_p^* / dr_1 = \partial L_p^d / \partial r_1 > 0.$$

Further, since  $\partial L_a^d / \partial r_2 > 0$ , it follows from (A9) that  $dw_c^* / dr_2 > 0$ . Since  $\partial L_p^d / \partial r_2 = 0$ , we have from (A7) that  $dL_p^* / dr_2 > 0$ .

## REFERENCES

- Agarwal, Bina, "Agricultural Mechanisation and Labour Use: A Disaggregated Approach," *International Labour Review*, January-February 1981, 120, 115-27.
- Bardhan, Kalpana, "Rural Employment, Wages and Labour Markets in India, A Survey of Research—III," *Economic and Political Weekly*, July 9, 1977, 12, 1101-18.
- Bardhan, Pranab K., (1979a) "Wages and Unemployment in a Poor Agrarian Economy: A Theoretical and Empirical Analysis," *Journal of Political Economy*, June 1979, 87, 479-500.
- , (1979b) "Labor Supply Functions in a Poor Agrarian Economy," *American Economic Review*, March 1979, 69, 73-83.
- , "Labor-Tying in a Poor Agrarian Economy: A Theoretical and Empirical Analysis," *Quarterly Journal of Economics*, August 1983, 98, 501-14.
- and Rudra, Ashok, "Terms and Conditions of Labor Contracts in Agriculture: Results of a Survey in West Bengal 1979," *Oxford Bulletin of Economics and Statistics*, February 1981, 43, 89-111.
- Basant, Rakesh, "Attached and Casual Labour Wage Rates," *Economic and Political Weekly*, March 1984, 19, 390-96.
- Bauer, Arnold J., "Chilean Rural Labor in the Nineteenth Century," *American History Review*, October 1971, 76, 1059-83.
- , *Chilean Rural Society from the Spanish Conquest to 1930*, Cambridge: Cambridge University Press, 1975.
- Bhalla, Sheila, "New Relations of Production in Haryana Agriculture," *Economic and Political Weekly*, March 27, 1976, 11, A23-30.
- Breman, Jan, *Patronage and Exploitation*, Berkeley: University of California Press, 1974.
- Eaton, B. Curtis and White, William D., "Agent Compensation and the Limits of Bonding," *Economic Inquiry*, July 1982, 20, 330-43.
- and ———, "The Economy of High Wages: An Agency Problem," *Economica*, May 1983, 50, 175-82.
- Eswaran, Mukesh and Kotwal, Ashok, "A Theory of Contractual Structure in Agriculture," *American Economic Review*, forthcoming 1985.
- Fogel, Robert W. and Engerman, Stanley L., *Time on the Cross*, Boston: Little, Brown, 1974.
- Hirashima, Shigemochi, *The Structure of Disparity in Developing Agriculture*, Tokyo: Institute of Developing Economies, 1978.
- Postan, M. M., *The Famulus: The Estate Labourer in the Twelfth and Thirteenth Centuries*, Suppl. No. 2, *Economic History Review*, 1954.
- Richards, Alan, "The Political Economy of Gutswirtschaft: A Comparative Analysis of East Elbian Germany, Egypt, and Chile," *Comparative Studies in Society and History*, October 1979, 21, 483-518.
- , *Egypt's Agricultural Development, 1800-1980*, Boulder: Westview Press, 1982.
- Rudra, Ashok, *Indian Agricultural Economics: Myths and Realities*, New Delhi: Allied Publishers, 1982.
- Sanghavi, Prafulla, *Surplus Manpower in Agriculture and Economic Development*, New Delhi; New York: Asia Publishing, 1969.
- Shapiro, Carl and Stiglitz, Joseph E., "Equilibrium Unemployment as a Worker Discipline Device," *American Economic Review*, June 1984, 74, 433-44.
- Smith, Thomas C., *The Agrarian Origins of Modern Japan*, Palo Alto: Stanford University Press, 1959.
- Stoft, Steven, "An Explanation of Involuntary Unemployment, Sticky Wages and Labor Market Structures," unpublished doctoral dissertation University of California-Berkeley, 1980.

# The Relative Inefficiency of Quotas: The Cheese Case

By JAMES E. ANDERSON\*

The nonequivalence of tariffs and quotas under a wide variety of circumstances is well established in economic theory. This study considers heterogeneity as a source of nonequivalence and demonstrates its practical importance. Where restriction is inevitable, the bias of economists in favor of price instruments is not thoroughly grounded; this study should help provide a foundation.

The point of departure from the equivalence model is differentiation of products. In practice, almost any quota system will be applied to what is in fact a commodity group. Legislators or chief executives restrict total imports of a group, possibly due either to lack of information or lack of a model which would allow a basis for derivation of a more detailed set of restrictions. Quota system administrators by default have wide authority to allocate trade within the group. Distribution of quota licenses in this case involves substantive resource allocation.

It is shown below that optimal allocation of quota licenses subject to a group total import constraint will produce uniform rent on use for each member of the group. This is equivalent to a uniform specific import tax save for revenue distribution. Administrators typically do not have competitive auctions in licenses that would accomplish the same end, but use simple rules such as base-year quantity or value proportions to hand out licenses that are type- and country-of-origin specific. Provided resale is frustrated, the allocation will be inefficient. Section I sets out the simple analytics of optimal and quantity proportion quota allocations and discusses factors influencing the size of inefficiency.

The purpose of this paper is to provide a case study of an import quota system with essentially the characteristics noted. Section II briefly discusses the U.S. dairy quota system. Section III reviews the econometric model of cheese demand used to calculate welfare effects. The results in Section IV demonstrate that the inefficiency loss is indeed substantial. The estimated AIDS model of the U.S. cheese quota system (see my 1983 paper) is used to calculate the welfare costs of the quota compared to an efficient tariff that produces the same aggregate quantity of cheese imports in the constrained categories. The analysis shows added cost amounting to over 15 percent of base expenditure on the constrained categories. Another revealing way of scaling the welfare cost compares the quota inefficiency loss with the total welfare loss from not having free trade. Switching from the quota to an efficient tax that yields the same quantity of imports gets us nearly one-third of the way back to free trade. Or, the quota distribution inefficiency adds nearly 50 percent to the unavoidable loss due to the constraint on imports.

Sections II and III discuss the considerable difficulties encountered in building an econometric model of the cheese quota system. The results of Section IV thus are subject to considerable bias, especially from aggregation, when regarded as a realistic account of the costs of the U.S. cheese quota system. This is unfortunate, but does not detract from the main point of this paper, which is to show a presumption that quota systems in practice are inefficient. The results of Section IV can be interpreted as plausible simulations of a quota system not unlike the U.S. cheese quota system. The size of the welfare loss together with the difficulties experienced in carrying on the study of quota incidence suggest that economists are on firm ground in opposing the trend toward use of quantity restrictions. Failing a return to price instruments, we should advocate data collec-

\*Professor of Economics, Boston College, Chestnut Hill, MA02167. This study was supported by NSF grant SES-7907085 to Boston College. Peter Dixon gave helpful comments, but should be held blameless for any remaining errors or obscurities.

tion in a manner which makes incidence analysis easier.

The results here are related to general propositions suggesting that aggregation seriously understates the deadweight loss of tax systems since aggregation reduces both estimated elasticities of substitution and estimated average taxes. See Peter Dixon (1978). The disaggregation of the model here avoids the type of bias Dixon studied, but, in addition, disaggregation by cheese type and countries uncovers a new source of loss worth over 15 percent of base expenditure on the restricted cheese categories. And even these results are based on a (forced) considerable aggregation of cheese types. Thus, painful as it is, disaggregation is strongly indicated for tax incidence problems.

### I. Efficient Quotas on Commodity Groups

Let us assume that political pressures result in an import quota to be applied to an aggregate commodity group. This takes the form of what Jagdish Bhagwati and T. N. Srinivasan (1983) call a "non-economic" constraint. Administration of the quota system is left to civil servants, who will typically use simple rules to allocate licenses under the constraint. While political interests build around the allocation systems, I regard these as second-order effects. They can be disregarded by an enlightened administrator who must nevertheless obey the legally mandated aggregate quota.

The optimal allocation is, intuitively, going to equalize rent on quota licenses across members of the group. This would be the equilibrium outcome of a competitive auction of licenses, and the inefficiency of any other allocation is seen to be a foregone arbitrage gain. The optimum can automatically be achieved by a uniform tax.<sup>1</sup>

<sup>1</sup>Formally, the analysis has much in common with that of Leslie Young and myself (1980). There the product differentiation was by states of nature. We analyzed optimal quota allocations across uncertain states subject to an average import constraint, and showed that the optimum is reached with a uniform tax. My article with Young (1982) considered constraints on other portions of the probability distribution of imports

Under alternative noneconomic constraints, a quota thus allocated "efficiently" can be inferior to, for example, an ad valorem tariff. The substitution toward higher value categories which reaps an arbitrage gain can also conflict with a noneconomic constraint on employment or revenue. See Robert Baldwin (1982) and Avinash Dixit (forthcoming). The issue of what constraint is appropriate is serious. It is always necessary to specify as closely as possible the deeper variable targeted by the political economic process and its relation to the trade control. Once this is done, ignorance of detailed elasticities and/or absence of a complete model will often necessitate an aggregate trade control.<sup>2</sup>

Consider the case of dairy quotas, for example. They arose out of a desire to limit domestic price support payments (see Section II below). This might imply a deeper noneconomic constraint on disturbing the subsidy budget by no more than a given amount. Possibly revenue raised on imports might also be relevant. Nevertheless it is reasonable to use an aggregate import constraint for two reasons. First, the stated objective of administrators was to place quotas on substitutable categories such that total annual imports of dairy products did not exceed a given milk equivalent tonnage, this being instrumental in limiting support payments to domestic producers (Section II). This effectively translates into a simple ag-

under uncertainty and we were able to rationalize tariff quotas, in which the tariff rate steps upward with import volume. In the present context of differentiated products, this translates into other constraints on distribution within the group besides the aggregate one. Again, analogous to our 1982 article, this will imply an optimum in which rents are equalized within subgroups. I should also note that everything said about quantitative restrictions and uniform specific taxes has immediate analogy with foreign exchange revenue restrictions and uniform ad valorem taxes. An inefficiently distributed quantity quota can, however, achieve higher welfare than an ad valorem tax in the achievement of a quantity constraint.

<sup>2</sup>This can have a variety of forms. For example, an efficiently applied constraint on foreign exchange value is indicated if employment is more closely linked to value than to quantity. Using the analysis above, this is readily seen to be equivalent to a uniform ad valorem tariff.

gregate quantity constraint. Presumably, administrators chose this objective because they were ignorant of demand elasticities when formulating quotas. By default all cheeses in a substitutable category were made identical. Second, optimizing the price support (possibly plus dairy import restriction revenue) subsystem is of dubious relevance when we observe that subsidy payments come from the general government revenue with little evidence of concern for magnitude as coverages change and when licenses are given away. Consideration of optimal fiscal systems, if it can be done sensibly at all, must be made a much bigger project. The simple noneconomic constraint is the natural primal analogue to partial equilibrium analysis.

Proceeding formally, I set out a simple analysis of a small trading country subject to a noneconomic constraint on aggregate imports within some group. The vector of import quantities in the constrained group is  $q = (q_1, \dots, q_m)$ . The vector of trade quantities in the other commodities is  $x$  (which contains positive elements for imports and negative elements for exports). Nontraded goods are irrelevant to the point I make, so are suppressed. External prices of  $q$  are given by the vector  $p$ , and external prices for the vector  $x$  are given by  $s$ . Allocation of production (if any in given categories) is irrelevant, so I assumed fixed endowments of goods that are produced;  $Z_1$  for constrained goods and  $Z_2$  for unconstrained goods.

The enlightened civil servant must choose a trade vector  $(q, x)$  to maximize utility of the representative consumer subject to the balance of trade constraint and the noneconomic constraint. Thus he solves the program:

$$(1) \quad \text{Max}_{q, x} U(Z_1 + q, Z_2 + x)$$

$$(\lambda) \quad p \cdot q + s \cdot x \leq 0$$

$$(\mu) \quad \iota \cdot q \leq \bar{q}$$

where  $Z_1, Z_2$  are constant production vectors,  $\mu$  and  $\lambda$  are Lagrange multipliers, and  $\iota$  is the vector of ones.

The first-order conditions are sufficient with concave utility, and yield in terms of a

numeraire from the unconstrained group:

$$(2) \quad (1/U_{2n})U_1 = p + (\mu/\lambda)\iota;$$

$$(1/U_{2n})U_2 = s,$$

where  $U_{2n} = \partial U / \partial x_n$  = marginal utility of the numeraire good,  $U_1 = \partial U / \partial q$ , and  $U_2 = \{\partial U / \partial x_i\}$ ,  $i \neq n$ .

The conditions (2) immediately imply that the unique global optimum can be decentralized with a uniform specific tax  $= \mu/\lambda$  on every element of  $q$ , the tariff proceeds being lump sum redistributed to the representative consumer. Equivalently, quota licenses can be competitively auctioned with redistribution of the proceeds.

In practice, quota licenses are usually distributed according to a simple rule, like quantity proportions in a base year. I now briefly consider theoretical factors affecting the magnitude of inefficiency created by such rules. With Leontief-type utility functions and endowments which grow radially (neutral growth), quantity rules will attain the optimal allocation in problem (1). Nonneutral growth of the  $Z$ s has an obvious effect. Less obviously, as the elasticity of substitution rises, two counteracting forces mediate the inefficiency of the rule. General analytical results are not attainable, but these forces are clear. First, the higher the elasticity of substitution, the less the harm associated with a given configuration of nonuniform taxes (equivalent to a quota distribution), since it is easier to substitute away from expensive categories. But second, the lower the elasticity of substitution, the less far the rule-given quantity restriction is from the optimal quantity change.

In simulation of the dairy quota system using a CES utility function and methods analogous to those of Section IV, the first effect predominated in a reasonable range of elasticity values (from 0.10 to 3), based on welfare loss relative to base expenditure. The second effect showed up when the inefficiency loss was scaled relative to the unavoidable inefficiency due to the constraint. This measure of relative inefficiency rose with the elasticity of substitution, but much more gently than the relative-to-base-expenditure measure fell. The latter typically dropped by



a factor of 10 in the range of elasticity values 0.10 to 1.0.

Similar exercises can be done for other rules or other utility functions. These results point to the sensitivity of empirical findings to the elasticities used, and thus emphasize the need for sound empirical work. The results of Section IV show large welfare losses; this and the theoretical analysis above indicate how easily a quantity rule can err badly.

## II. The U.S. Dairy Industry and the Quota System

I first consider the institutional and legal structure of dairy import controls and then the resulting market structure assumed. I aim to justify my use of 1) the simple form of "noneconomic" import constraint, and 2) a partial equilibrium model.

The U.S. dairy quota system originated as a by-product of the dairy price support price system. The main element is support of a manufacturing grade milk price, but there are also butter and American-type cheese support prices. At various times, beginning in 1937, quotas have been imposed on a variety of cheeses, as well as butter, nonfat dry milk, substances high in nonfat dry milk or butterfat, and other milk products. The president has very broad authority to proclaim, raise, or suspend these trade controls,<sup>3</sup> but in practice revision has been limited.

<sup>3</sup>Presidential authority for the quota is drawn from Section 22 of the Agricultural Adjustment Act of 1933 (originally added in 1935 and subsequently revised extensively). It directs the Secretary of Agriculture to advise the president whenever he has reason to believe any article is imported in such conditions and such quantities as to: 1) render or tend to render inefficient or materially interfere with any price support or stabilization program relating to agricultural commodities, or 2) reduce substantially the amount of any product processed in the United States from any agricultural commodity with respect to which any such program is being undertaken. If the president agrees there is reason for such belief, he directs the U.S. International Trade Commission to conduct an investigation and submit a report. The president is authorized, based on such findings, to impose such fees or quotas in addition to the basic duty as he shall determine necessary. The president may designate the affected article by physical qualities, value, use, or upon such other basis as he shall determine. The president may revise or suspend all previously proclaimed fees or quotas. Any decision by the president as to the facts under this authority is final.

In the timeless model of Section I, the quota allocation problem is faced once and for all. In practice, the set of dairy quotas reflect an evolution of broadening the scope of coverage as more products appear to be sufficiently substitutable with domestic products to warrant control. Thus the date of inception and benchmark period (upon which proportionate allocations are based) differ for different cheeses and other products. The spirit of the quota system is, however, to set detailed levels of imports on a basis consistent with historic proportions. The presidential authority does not stipulate that this principle is to be carried through to allocation by country of origin, but the administrators have so proceeded. An important implication of the evolving nature of the quota system is that the administrators help create the marginal changes in the system issued by the president. Significantly, they appear to regard their objective as a target level of milk equivalent tonnage in groups identified as substituting for domestic milk products. (See, for example, Harlan Emery, 1969, pp. 9-11.) Thus both the history of quota setting and the stated objective of the quota administrators supports a noneconomic constraint of the simple form in (1). I assume that any second-order political interests do not create binding constraints.

The current allocation system, on the other hand, does apparently create *economically* binding constraints. The USDA quota system administrators develop license allocation by commodity by country from base-year allocations in the legally mandated categories. They claim to have effective auditors who implicitly frustrate resale of licenses. Measured differences in average quota rent margins, reported below, bear the claim out. This creates the basis for my study of efficient alternative dairy product import controls.

Butter and other milk product quotas are ignored here on the grounds that they are rather simple undifferentiated products with low elasticity of substitution with cheese or with each other. On prior reasoning, the quota allocations in these categories might be reasonably appropriate. The group of cheese controls appears to represent a much closer approximation to the theoretical model

above. Cheese is a highly differentiated product with fairly high within-group elasticities of substitution. The Census trade data have nine commodity categories of continuously imported cheese plus a catchall, with six of the nine plus a part of the catchall category subject in part or whole to quota constraint. This creates the circumstances for a case study of quota inefficiency when we convert the quota constraints into an overall constraint on the six categories. Inefficiency will arise due to 1) inefficient allocation by country within the same cheese type (a matter of administrative discretion), and 2) inefficient allocation over types (partly mandated in current presidential proclamations, but presumably easily changed within the spirit of the presidents' previous exercises of authority).

I now rationalize a partial equilibrium approach to the market for imported cheese. The United States is a small consumer of foreign cheese, hence foreign price of foreign cheese is reasonably taken to be exogenous. Domestic price in quota constrained categories is of course endogenous. The U.S. consumption of cheese is a tiny fraction of total food consumption, so ignoring spillover effects onto noncheese prices may be reasonable. Domestic price of domestic cheese is endogenous save when the government is maintaining a floor price through large purchases. Endogenous prices are fitted to an implicit reduced form in the econometric work reported below.

In principle, it is possible to model the supply side of the domestic cheese industry. Domestic cheese production and sale is dominated by the government's milk, cheese, and butter price supports, but not entirely so. Successful dealing with these markets probably requires a model of how the government sets its supports. The highly politicized process attending this in the 1970's is common knowledge. An attempt to deal with modeling domestic cheese supply was dropped at an early stage because in the welfare (as opposed to econometric) analysis it is permissible to treat the U.S. price of U.S. cheese and related dairy products as exogenous to import policy changes, being separately set by the government and main-

tained by price support purchases. This is effectively true in a large part of the sample (1964-79), particularly in the later years. The displacement effect of foreign price changes on domestic cheese categories can be calculated using the demand functions for imported cheese and relying on the symmetry implied by demand theory. The implied change in price support purchases of domestic cheese is carried through as part of the analysis.

I now consider briefly problems of aggregation. With the data available it proved to be impossible to match exactly the quota constraints with the commodity categories. My solution was to aggregate. Quota authority lies with the USDA, with different statistical codes from the Customs (TSUS) code, which are different still from the Census (SITC) codes. Country-level quota information was available, but attempts to build a complete model of import supply were frustrated by aggregation problems and complexities in how quota licenses are used. (See the Appendix to my earlier paper for more detail.) The econometric model used in the next sections glosses over quota system complications by assuming that for each of the six categories constrained by quota, the entire set of members is treated uniformly. Foreign price of any cheese is exogenous, but domestic price of domestic cheese and all quota-constrained cheese is endogenous and fitted to an implicit reduced form. This creates aggregation bias of unknown sign and magnitude. Similarly, the welfare analysis is biased as a "true" account of the cost of U.S. quota system inefficiency.

### III. The Model of U.S. Imported Cheese Markets

The preceding section essentially justifies a partial equilibrium approach to modeling the effect of switching from quotas to tariffs. The U.S. domestic cheese prices are assumed to be made exogenous by government price support, other food prices are not affected, and the United States is a small customer in world dairy markets, hence foreign prices are exogenous. For the welfare evaluation, the endogenous variables are assigned their sample mean values in the quota regime.

For econometric modeling, U.S. domestic cheese prices are made endogenous (since they are at least some of the time) and the U.S. price of quota constrained foreign cheese is also endogenous. I thus evade modeling domestic supply of both foreign and domestic cheese except as it is embedded in an implicit reduced form. We do require a set of demand functions for imported cheese.

A nine-equation model of imported cheese demand is estimated for the years 1964–79. A food and beverage expenditure function is assumed to exist for a representative U.S. consumer identifiable in the aggregate data. (See my earlier paper for discussion of the implied assumptions.) The AIDS expenditure function is used, with the further wrinkle that only a tiny portion of total food and beverage consumption is estimated, that for nine imported cheeses. See Angus Deaton and John Muellbauer (1980a,b) for a general treatment of AIDS. The AIDS function is used because it is a flexible functional form with particularly simple capability for allowing nonhomothetic preferences while permitting exact linear aggregation.

The AIDS expenditure function is defined as

$$(3) \log e(p_1, \dots, p_n, u) = \sum_{i=1}^n \alpha_i \log p_i + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \gamma_{ij} \log p_i \log p_j + u\beta_0 \prod_{j=1}^n p_j^{\beta_j},$$

where  $u$  = utility indicator,  $e(\ )$  = expenditure function, and  $p_i$  = price of good  $i$ .

By Shephard's Lemma, the compensated demand functions are  $\partial e / \partial p_i = c_i(p_1, \dots, p_n, u)$ . It is more convenient to place these in compensated expenditure share form:  $w_i = (p_i/e)(\partial e / \partial p_i)$ , where  $w_i$  is the share of expenditure on  $i$ . Next, (3) can be solved for  $u$  using  $e = y$ , expenditure = income. The resulting indirect utility substituted into the compensated share equations yields the uncompensated expenditure share equations:

$$(4) w_i = \alpha_i + \sum_{j=1}^n \gamma_{ij} \log p_j + \beta_i \log(y/P),$$

where  $P$  is the true cost of living index

defined by

$$(5) \log P = \sum_j \alpha_j \log p_j + \frac{1}{2} \sum_i \sum_j \log p_i \log p_j.$$

Deaton-Muellbauer argue that  $P$  is usually closely approximated by standard indices. The restrictions implied by consumer theory are

$$(6) \gamma_{ij} = \gamma_{ji}, \quad \text{symmetry;}$$

$$(7) \sum_{i=1}^n \alpha_i = 1, \quad \sum_{j=1}^n \gamma_{ij} = 0, \quad \sum_{j=1}^n \beta_j = 0, \quad \text{homogeneity;}$$

$$(8) \{k_{ij}\} = \{-\delta_{ij}w_i + w_iw_j + \beta_i\beta_j \log y/P\}, \quad \text{negative semidefinite, concavity.}$$

In (8),  $\delta_{ij}$  is the Kronecker delta. The matrix  $k = \{k_{ij}\}$  is a simple transform of the Slutsky matrix, with the same properties.

Equation (4) in its stochastic form is fitted to 1964–79 U.S. quarterly data. A nine-equation set of  $w_i$ s is fitted to the logs of a set of fifteen cheese prices (including domestic types), a general price index for food, and a "real food and beverages" expenditure term  $y/P$ . It is convenient to normalize the price data by division by the arithmetic mean. The transformed log price data is then all zero at the point of arithmetic means. In the base case estimation, symmetry and homogeneity are imposed on (4). Concavity (for the nine imported cheese prices) is imposed at the point of means and holds in small region about it as discussed in my earlier paper. The results are reasonably "good." The main sources of possible specification error are in aggregation over consumers and over cheeses. The data do not permit an attack on these problems. To give some feel for the demand structure that results, Table 1 presents point estimates of compensated price elasticities. They are defined by

$$(9) e_{ij} = -\delta_{ij} + w_j + \frac{\gamma_{ij}}{w_i} + \frac{\beta_i\beta_j \log y/P}{w_i},$$

TABLE 1—PRICE ELASTICITIES AT POINT OF MEANS

	Blue	Cheddar	Edam-Gouda	Roquefort	Romano, etc.	Swiss	Gruyere	Grating Pecorino	Nongrating Pecorino
Blue	-1.22567044	-0.36196737	0.00008190	0.00002466	0.29799052	-1.46954474	0.00008878	-0.19759316	0.00006180
Cheddar	-0.12363571	-6.38763977	-0.24473892	-0.07932095	-0.38251052	0.00027011	0.00008878	0.00013060	0.09562920
Edam-Gouda	0.00003209	-0.28271565	-1.77124995	0.00002466	0.00008057	0.00027011	-0.94083556	0.00013060	0.03080344
Roquefort	0.00003209	-0.41988079	0.00008190	-0.84767953	0.04624139	0.00027011	0.05061753	0.70071737	-0.53713195
Romano, etc.	0.11871424	-0.44639119	0.00008190	0.01020211	-1.42083450	0.00027011	0.00008878	0.00013060	0.10335093
Swiss	-0.17431817	0.00009477	0.00008190	0.00002466	0.00008057	-2.03426212	-0.73051634	0.07559175	0.00006180
Gruyere	0.00003209	0.00009477	-0.86768188	0.01019569	0.00008057	-2.23981415	-1.70398023	0.54818688	0.00006180
Grating Pecorino	-0.04844285	0.00009477	0.00008190	0.09507469	0.00008057	0.15623685	0.36952309	-1.01506819	0.60988780
Nongrating Pecorino	0.00003209	0.14574259	0.04064101	-0.15465579	0.13498088	0.00027011	0.00008878	1.29433575	-1.23172324

where  $\delta_{ij}$  is the Kronecker *delta*. The elasticities are evaluated at the point of means of the shares. A reader skeptical of the econometric methods used is free to regard these and the underlying parameters as plausible guesses to be used in simulation.

The estimated imported cheese demand functions and the symmetry restrictions of the expenditure function ( $\gamma_{ij} = \gamma_{ji}$ ) permit a complete account of the effect of a change in import policy. All relevant parameters of the expenditure function are known, as are the relevant cheese demand parameters. The analysis of Section I derives the efficient policy as being a uniform tax that would achieve the same aggregate quantity of constrained cheese. The estimated demand functions evaluated at the point of means can be used to solve for the efficient tax. A welfare analysis can then be carried through, based on the consequences of the change in the vector of constrained imported cheese prices.

#### IV. The Welfare Loss of Inefficient Dairy Quotas

The model used to evaluate the welfare effects of the change from quotas to the efficient tax system is standard. The changes are measured in 1) the representative (hence aggregate) consumer's expenditure,<sup>4</sup> 2) the

government's tax receipts, and 3) the quota license holder's loss of rent. Ordinarily, we would need to add in the change in income from production, but it is assumed domestic product prices are constant. For noncheese products, this is because it is assumed the changes in imported cheese prices have negligible effect on demand. For all domestically produced cheese, it is assumed for convenience that the government supports the price by purchases.

In practice government price support is given only for American-type cheese. The effect of assuming support for non-American-type cheese is to wash out some second-order effects on welfare via domestic product price changes. These have negligible first-order effects, since with supply equal to demand, the expenditure effect and production revenue effect cancel. I make this assumption to avoid what might be an inappropriate use of the implicit model of domestic supply. I report below on both American-type and total support payment changes.

Formally, the expenditure function may be written as  $e(p, s, u)$ , where  $p$  is the vector of domestic prices for constrained goods,  $s$  is the vector of domestic prices for unconstrained goods, and  $u$  is the parameter utility level. The vector  $s$  is made parametric either by: (a) the small country assumption for traded goods other than cheese, (b) standard partial equilibrium "smallness" arguments

<sup>4</sup>A superior alternative is available. Equation (3) can be solved for  $u$  with  $e = y$ , with  $u$  now indirect utility. The result can be substituted back into (3), which now becomes the (log of) income compensation function, or (log of) the money metric of utility change,  $\log e(p^0, u(p, y))$ , where  $p^0$  is a base price vector and  $p$  is an alternative price vector. This is the ideal welfare measure. For the case studied, movement of  $u$  is

guaranteed to be trivial and it is simpler to keep it constant. Thus we use the expenditure function (3) as the basis of welfare calculations (which corresponds to use of compensating rather than equivalent variations).

with respect to nontraded goods, or (c) government floor price support purchases for domestic cheese products displaced by import price changes. The vector  $p$  is built up from foreign prices:

$$(10) \quad p = p^* + (\text{margins}) + t,$$

where  $p^*$  is the parametric foreign price vector,  $(\text{margins})$  is the assumed parametric vector of distribution margins, and  $t$  is the tax-cum-quota-rent vector. If  $p^0$  and  $p^1$  are two vectors created by alternative restriction policies, the change in expenditure is

$$(11) \quad \Delta e = e(p^1, s, u) - e(p^0, s, u).$$

The change in quota rent is simply the original demand times the quota rent:  $\partial e / \partial p \cdot t$  in the case where there was no initial tax on imports. In practice, there is in fact a portion of  $t$  which is a tax,  $t_g$ , so the actual rent lost is

$$(12) \quad \Delta \text{rent} = \frac{\partial e(p^0, s, u)}{\partial p} \cdot (t - t_g).$$

Now I account for government support payments. The displaced domestic demand is replaced by government purchases. Let  $s_d$  denote the supported products price vector, where  $s_d$  is composed of some element of  $s$ . The change in support payments

$$(13) \quad \Delta \text{support} = \left[ \frac{\partial e}{\partial s_d}(p^1, s, u) - \frac{\partial e}{\partial s_d}(p^0, s, u) \right] \cdot s_d,$$

where  $\partial e / \partial s_d$  is the compensated demand vector for supported products. Below I present two measures for (13). One is the American-type cheese support payments—the only cheese so supported in reality. The other is the assumed support payments for all domestic cheese in the quota-protected categories.

The first variant on (13) is a lower bound for actual support payment changes, since additional small changes in demand for other supported dairy products might be expected.

The second minus the first is an upper bound for additional welfare loss, due to changes in other domestic cheese prices, since cheeses other than American-type are not supported. It should ordinarily far exceed the actual small secondary effect we are neglecting.

Finally, I account for changes in tariff revenue. On the group of constrained cheeses, the revenue is based on the common scalar efficient tax  $t^*$  solved for to satisfy the constraint:

$$(14) \quad \frac{\partial e}{\partial p}(p^* + (\text{margins}) + t^* \iota, s, u) \cdot \iota = \bar{Q},$$

where  $\iota$  is the vector of ones. Revenue is collected in the amount  $t^* \cdot \bar{Q}$ . From this must be deducted the revenue previously collected by the government with its taxes on this group of cheese:  $\partial e / \partial p \cdot t_g$ , where  $t_g$  is the initial specific tax vector (converted from the ad valorem actual tax at the point of mean prices). Thus the change in revenue is

$$(15) \quad \Delta \text{revenue} = t^* \bar{Q} - \frac{\partial e}{\partial p}(p^0, s, u) \cdot t_g.$$

Note that (15) neglects the change in tariff revenue on the non-quota-constrained imported cheeses. This is done on the principle that the government budget constraint is supposed to be irrelevant, although I have kept track of  $\Delta \text{support}$ .<sup>5</sup> The amounts involved are small, in any case.

I report two welfare measures below. Gross welfare is the logically consistent measure corresponding to the theoretical model. It neglects  $\Delta \text{support}$ .

<sup>5</sup> If government revenue is made to matter, a revenue constraint is added to the problem, and in the optimal solution a Ramsay price element is added to the uniform part of the optimal tax, as noted in Section I. It should be noted that another version of the problem kept the existing ad valorem taxes as given politically, and optimized the quota system analogously to the procedure above. This procedure worked nearly as well as the solution reported. This is not surprising in view of the low level and low dispersion of the ad valorem tariffs.

(16)  $\delta$  gross welfare

$$= -\Delta e - \Delta \text{rent} + \Delta \text{revenue}.$$

Net welfare subtracts from this the change in support payments on American-type cheese.

The efficient tax and its welfare consequences in (12)–(15) are calculated using the parameter estimates detailed in my 1983 paper. The procedure can be performed at any data point. I restrict attention primarily to the point of sample means, but discuss briefly the evaluation of extremal values. In the estimated AIDS system (4), the income terms  $\beta_i$  are effectively zero, so that we may impose homothetic preferences for evaluation purposes.<sup>6</sup> The uncompensated share equations (4) are equal in this case to the compensated share equations. I solve them for quantity demanded in the six quota-constrained categories. The government's noneconomic constraint is to hold aggregate imports across the six cheeses to the given sample mean quantity. We know the efficient solution requires a uniform specific tax  $t$ . The new domestic price for import  $i$  will be  $\hat{p}_i + t^*$ , where  $\hat{p}_i$  is the margin augmented foreign price, or free-trade price. I solve for the optimal tax  $t^*$  in

$$(17) \quad \bar{Q} = y \sum_{i \in QR} 1/(\hat{p}_i + t) \\ \times \left\{ \alpha_i + \sum_{k \in QR} \gamma_{ik} \log((\hat{p}_k + t)/p_k^0) \right\}.$$

The term in curly brackets is the new expenditure share for import  $i$ ,  $i \in QR$ . The  $QR$  is the index set for constrained cheeses, and  $\bar{Q}$  is the given sample mean aggregate quantity. Operations outside the curly brackets convert shares to quantities, and aggregate. Inside the brackets,  $p_k^0$  is the sample mean (quota regime) price. Its presence is necessitated by the econometric procedure of using sample-mean-scaled prices. The great convenience of this method is that other

arguments of (4) are zeroed out in evaluation at the sample mean (all right-hand side data are scaled by the sample mean and  $\log(x/\bar{x}) = 0$  at  $x = \bar{x}$ ).

Equation (17) is well behaved as a function of  $t$  at the sample means point, and Newton's method converged very quickly. For evaluation away from the point of means, (17) is somewhat more complex in form, and for extremal values of  $p_k$  has possible multiple roots (one case was discovered). This difficulty is related to the problem of non-concavity of the estimated  $e(p, u)$  at extreme values of  $p$ .<sup>7</sup> I restrict primary attention to the point of means for this reason, but report results at other data points because the extremal price distortions are precisely where the quota inefficiency ought to be greatest.

From evaluation of (17), the resulting new domestic price vector  $\{\hat{p}_i + t^*\}$  for six categories is then plugged into the expenditure function (3). The relative change in expenditure is obtained by subtracting the log expenditure at the initial point from that at the new point, then exponentiating. The absolute size is obtained by multiplying by base expenditure. Making use again of the zeroing-out property of evaluation at the point of means, I report the application of (11):

$$(18) \quad \Delta e = \exp \left[ y \sum_{i \in QR} \left\{ \alpha_i \log((\hat{p}_i + t^*)/p_i^0) \right. \right. \\ \left. \left. + \frac{1}{2} \sum_{k \in QR} \gamma_{ik} \log((\hat{p}_i + t^*)/p_i^0) \right. \right. \\ \left. \left. \times \log((\hat{p}_i + t^*)/p_k^0) \right\} \right].$$

<sup>7</sup>The efficient quota problem (1) has a unique global optimum defined by the equations (2), the first-order conditions. Equation (17) is the sum of the rows of the inverse of system (2), hence need not yield a unique  $t^*$ . Among its solutions must lie the optimal tax. Note that the derivative of (17) with respect to  $t$  is ordinarily negative, the more so as own-effects predominate. "Good behavior" has the derivative negative in the relevant range, thus yielding a unique solution. Nonconcavity evidently exacerbates the "bad behavior" problem, since it is associated with prominent cross effects.

<sup>6</sup>One income elasticity differs statistically significantly from unity in the third place to the right of the decimal point.

Government tax revenue is lost by doing away with the existing ad valorem taxes, and gained by the new specific tax. The new revenue for the six categories is  $t^* \cdot \bar{Q}$ . The revenue lost is subtracted from this, so the application of (15) is

(19)  $\Delta \text{revenue}$

$$= t^* \bar{Q} - y \sum_{i \in QR} \frac{\tau_i}{1 + \tau_i + \text{cif}_i + \text{whl}} \alpha_i,$$

where  $\tau_i$  = ad valorem tariff rate on  $i$ , and other variables are cif and wholesale margins, respectively. In (19) I make use of evaluation at the point of means, so  $\alpha_i$  is the relevant share of imported cheese. I divide by  $1 +$  the tax rate plus the cif percent markup (cif) + the wholesale percent markup (whl) to convert the domestic value in  $\alpha_i$  into a foreign f.o.b. value.

I calculate  $\Delta \text{support}$ , the application of (13), using evaluation at the simple mean, as

(20)  $\Delta \text{support}$

$$= y \sum_{j \in S} \sum_{k \in QR} \gamma_{jk} \log((\hat{p}_k + t^*)/p_k^0),$$

where  $S$  is the set of supported cheeses and other variables are previously defined.

Below, the gross welfare measures represent the pure logic of quota inefficiency measurement; the net measures have increases in cheddar support payments subtracted out. These overestimate the change the government would actually experience, since aged cheddar is not supported. I also report a support payment change on the assumption that *all* domestic cheese displaced must be purchased by the government.

The net welfare figure presented below is biased very far downward, the gross welfare figure is biased somewhat upward by its failure to account for the displaced domestic cheese not subject to support payments. Presumably this will cause changes in the price of domestic cheese, the welfare consequences of which will have to be traced through. For the efficient tax solutions the bias is very minor, since the displacement is under 3 percent of the total market in each cheese.

For the free-trade solutions, the case is more serious (displacement of 9 and 19 percent in the two cases below), and for cheddar in particular the bias will be substantial (displacement of 19 and 39 percent). In the absence of a model of domestic cheese production there was little to be done about this. It was not significant to the study, since the free-trade solutions are produced primarily to scale the efficient tax solutions of interest.

The final category in the welfare calculation is the loss of quota rent. The entire rent change is treated as a loss, though some rent probably goes to foreigners or is dissipated in rent-seeking activity.

Two cases are detailed below. In one, the allocation of country imports does not shift to low-cost producers; in effect there is only one representative exporter for each cheese. The average base-period f.o.b. unit value is thus used to form  $\hat{p}_i$ . The other case uses a conservatively picked lower-cost producer as the importer. The new import price was far from the lowest and came from a supplier with deep markets and substantial share already (for example, New Zealand or West Germany). The new lower  $\hat{p}_i$  then serves as the base for recalculating all previous steps. This procedure inevitably has considerable danger of (dis-)aggregation bias. For the composite SITC 0240025, Romano-Parmesan-Provolone, etc., no such reallocation is allowed. For the other categories, the gain in realism appears worthwhile. For one category, Swiss, I have domestic price series for Finnish, Austrian, and Swiss origin. The latter commands about a 10 percent premium. The cif import unit values in these categories diverge by 30 percent or more. Thus there is clearly a country reallocation gain to be reaped.

As a basis for the commodity reallocation gains, Table 2 presents the sample mean ad valorem tariff rates and the mean ad valorem equivalent quota margins for the six constrained categories. The sample mean foreign port values of the various imported cheeses range from .6 to 1.6, so Table 2 obviously implies an inefficient specific tax equivalent. At most data points, the quota margins show *much* wider variation over categories than at the point of means shown in

TABLE 2—MEAN POLICY DISTORTIONS ON  
IMPORTED CHEESE, 1964–79

	Tariff	Quota Ad Valorem Equivalent <sup>a</sup>
Blue	.15	.025
Cheddar	.15	.33 <sup>b</sup>
Edam-Gouda	.15	.33
Romano, etc.	.20	.14
Swiss	.11	.20
Gruyere	.11	.33 <sup>b</sup>

<sup>a</sup>Aggregating biases these estimates downward.<sup>b</sup>No domestic price of imported cheese available. Estimate constructed by applying the mean of markups of a large number of components from other categories.

the table, so greater efficiency gains can be expected.

The results of efficient allocation are in Tables 3 and 4. Table 3 has the optimal specific tax (the solution of (17)), and the unscaled amounts (in millions of base dollars). A negative sign is a fall in the category whether increasing or decreasing welfare. Since imported cheese is small relative to *GNP* or even advanced war equipment, the numbers in Table 3 are not impressive. They are supplied mainly to allow alternative treatment by the reader of dubious categories like support and rent.

Scaling the absolute magnitude of changes is critical to seeing their importance. One scale used is base expenditure on the six imported constrained cheeses. The other scale used is the welfare change implied by going to free trade (redo the calculations above with  $t = 0$ ).

Two appropriate scales are in Table 4. The first four columns express the savings as a proportion of base expenditure on the six cheeses. Columns 2 and 4 contain the net and gross welfare proportionate changes implied by free trade. The percentages in columns 1 and 2 are fairly impressive. My best guess is that at the point of means, an efficient tax saves us over 15 percent of base expenditure (row 2, column 1 of Table 3). Standard triangle welfare loss measurement seldom turns up relative gains of more than 1 percent.

Perhaps the most revealing measure of relative inefficiency of the quota is in column 5,

the result of dividing column 1 by column 2. For the two cases shown, the number says that efficient taxation takes us 10 and 30 percent, respectively, of the way back to free trade.

Table 5 contains percentage changes in imported cheese consumption in the efficient tax solutions. Short of implementing an efficient tax, the numbers in Table 5 could be used to guide a reallocation of quota licenses. Quota administrators might in fact not be surprised at the numbers in Table 5, since they are linked to the size of current sample mean quota margins and tariffs. Compare Tables 5 and 2. In the more extreme data points, the quota margin variations cause less obvious substitution patterns to be optimal.

The most disappointing result of Tables 2 and 3 is the small gross welfare gain derived by going to the efficient tax without supply reallocation. This is the pure-differentiated-product" effect (though the gain from country reallocation operates partly through the differentiated product effect). Only about 1/4 of 1 percent of base expenditure is saved by efficient allocation. If evaluation is done at other data points, however, this figure rises to over 2 percent at the maximum. The mean of all gross welfare efficient tax gains (as a percent of base expenditure) is .0044, with a standard deviation of .0043. Over 10 percent of the evaluations yield gross welfare savings in excess of 1 percent of base expenditure. Interestingly, extreme data points also frequently change the sign of net welfare changes. The reason is that support payments often drop rather than rise due to substitution effects. The mean net welfare change as a percent of base expenditure is .012 with a standard deviation of .14. The maximal value of the relative net welfare change is .41 and the minimal value  $-.39$ . More could be made of these extreme values were it not for doubts about their reliability due to the estimated demand system being moved outside the reasonable approximation region.

This brings up the important general issue of sensitivity of results to the estimation of the demand system, especially to the "concavifying" restrictions. As discussed in my earlier paper, concavity was imposed by set-



TABLE 3—EFFECT OF SWITCH TO EFFICIENT TAX ON CONSTRAINED CHEESE, SAMPLE MEAN POINTS (in millions of dollars)

	Tax (\$/lb)	Expen- diture (a)	Rent (b)	Tax Revenue (c)	American Support (d)	All Support (e)	Gross Welfare (f)	Net Welfare (g)
No Supply Reallo- cation	.224	-.942	15.52	14.98	25.14	27.45	.402	-24.74
Supply Reallo- cation	.405	-.80	15.52	36.3	51.95	392.13	21.58	-30.37

Note: Col. (f) =  $-(a)-(b)+(c)$ ; Col. (g) =  $-(a)-(b)+(c)-(d)$ .

TABLE 4—SCALED WELFARE CHANGES OF EFFICIENT TAX AND FREE TRADE IN CHEESE

	Gross Welfare/ Base Expenditure		Net Welfare/ Base Expenditure		Relative Inefficiency
	Efficient Tax (1)	Free Trade (2)	Efficient Tax (3)	Free Trade (4)	
No Supply Reallocation	.0028	.087	-.17	-1.70	.103
Supply Reallocation	.152	.497	-.21	-3.27	.306

Note: Col. (5) = (1)/(2).

TABLE 5—CHANGES IN CHEESE CONSUMPTION IN EFFICIENT TAX SOLUTION

	No Supply Reallocation (1)	Supply Reallocation (2)	Base Food Expenditure Share $\times 10^9$ (3)
Blue	-.123	-.126	32,092
Cheddar	.292	.619	94,767
Edam-Gouda	.216	.474	81,898
Romano, etc.	-.022	-.159	80,587
Swiss	-.023	-.067	270,110
Gruyere	.120	.158	88,780

Note: Percent changes in consumption.

ting linear constraints which were incrementally tightened. Tables like 3 and 4 were generated automatically for a large number of concavifying restriction values. The numbers in especially the first 3 rows of Table 4 were highly robust with respect to these changes. It is notable that this statement

applies even to evaluations where concavity at the point of means was not attained. Another check on sensitivity is afforded by using the Cobb-Douglas limiting case of the AIDS. The Cobb-Douglas parameters are the mean shares. It implies price elasticities which differ substantially from the point estimates of AIDS elasticities at the point of means. Compare the 9 rows and columns of Table 1 with minus the identity matrix, which closely approximates the compensated Cobb-Douglas price elasticity matrix given the smallness of shares. Net and gross welfare changes are approximately equal for the Cobb-Douglas, so we check the sensitivity of gross welfare changes to the functional form. When analogous calculations are carried through to produce the upper left cell of Table 4, gross welfare inefficiency relative to base expenditure, with no supply reallocation, the corresponding number is .0013, versus the table's .0028. With reallocation,

the Cobb-Douglas gain is .147 of base expenditure vs. .152 in Table 4. The data reject the Cobb-Douglas model very strongly (the equation  $R^2$ s in the AIDS average over .50). Nevertheless, the Cobb-Douglas is so familiar and easy to use that it makes a convenient benchmark. A bias factor so small when even so gross a misfit as the Cobb-Douglas is imposed suggests that the underlying inefficiency is indeed substantially accurately measured by the methods of this study at the point of means. Something like this relationship is maintained even at more extreme values of the data.

The impressive magnitudes in Tables 4 and 5 validate the main proposition of this study. Quota systems on heterogeneous commodities are likely to add substantially to the inefficiency of protection. Subsidiarily, the results bear out claims that any welfare loss measure will be greater the greater the disaggregation. Painful as it may be, commodity detail is strongly indicated for studies of protection. Finally, the results together with difficulties experienced in achieving comparability of data argues strongly that government statistical bureaus should collect data on quota categories automatically in the form required to assess incidence.

#### REFERENCES

- Anderson, James E., "An Econometric Model of Imported Cheese Demand," mimeo., 1983.
- \_\_\_\_\_, and Young, Leslie, "The Optimality of Tariff-Quotas," *Journal of International Economics*, November 1982, 13, 337-52.
- Baldwin, Robert E., "The Inefficiency of Trade Policy," *Essays in International Finance*, No. 150, Princeton University, 1982.
- Bhagwati, Jagdish N. and Srinivasan, T. N., *Lectures on International Trade*, Cambridge: MIT Press, 1983, ch. 24.
- Deaton, Angus, and Muellbauer, John, (1980a) *Economic Theory and Consumer Behavior*, Cambridge: Cambridge University Press, 1980.
- \_\_\_\_\_, and \_\_\_\_\_, (1980b) "An Almost Ideal Demand System," *American Economic Review*, June 1980, 70, 312-26.
- Dixit, Avinash, "Tax Policy in Open Economies," in Alan Auerbach and Martin Feldstein, eds., *Handbook of Public Economics*, Amsterdam: North-Holland, forthcoming.
- Dixon, Peter, "Economies of Scale, Commodity Disaggregation, and the Costs of Protection," *Australian Economic Papers*, June 1978, 70, 312-26.
- Emery, Harlan, "Dairy Price Support and Related Programs, 1949-1968," Agricultural Economic Report No. 165, U.S. Department of Agriculture, Washington, July 1969.
- Young, Leslie and Anderson, James E., "The Optimal Policies for Restricting Trade Under Uncertainty," *Review of Economic Studies*, November 1980, 46, 927-32.

# Testing for the Effectiveness of Wage-Price Controls: An Application to the Carter Program

By JOHN B. HAGENS AND R. ROBERT RUSSELL\*

During the last twenty-two years, income policies have been in effect in the United States more than half the time. The Kennedy-Johnson wage-price guideposts lasted almost six years, although they were tattered and torn by the time they were abandoned in 1967. The Nixon Administration's Economic Stabilization Program (ESP)—the only peacetime mandatory controls program in the United States—lasted four years, although the second half of the program (Phases III and IV) was a period of decontrol. The Carter Administration's Pay and Price Standards Program lasted a little more than two years, although the second year was one of thinly disguised decontrol.

It is not unlikely that another incomes policy will find its way into a government anti-inflation policy package sometime in this decade. The popular press regularly runs editorials and guest articles on the need for an incomes policy, and some presidential aspirants have incorporated incomes policies (especially tax incentive programs) into their informal platforms.

Unfortunately, the economics profession offers little guidance to policymakers about the efficacy of such programs. Diametrically opposite conclusions are reached by arm-chair reasoning (which is often predicated on strongly held ideological preconceptions), and the results of empirical tests too often are conflicting, ambiguous, or inconclusive.

These conflicting empirical results are partly attributable to the notorious non-robustness of wage and price equations (and

of other models of inflation, such as extreme-monetarist and ARIMA models),<sup>1</sup> but part of the problem is the paucity of thought that has gone into the formulation of the hypothesis to be tested. The standard approach is a test for the statistical significance of the estimated coefficient of an "on-off" intercept dummy variable. There are, however, many ways in which an incomes policy might conceivably affect the structure of an inflation model. Articulation of these different possible effects can be found in policymakers' justifications of particular incomes policies—for example, bringing wage demands into line with productivity growth, shifting the (short-run) Phillips curve downward, changing the tradeoff between unemployment and inflation (i.e., changing the slope of the short-run Phillips curve), retarding inflationary expectations, breaking the wage-price or the wage-wage spiral, and preventing extraneous shocks from getting built into the wage-price process. These alternative versions of the objectives of incomes policies have profoundly different implications for the effect on the structure of a model of the inflation process. Yet virtually no attention has been paid to the diversity of potential effects of such policies.

The purpose of this paper is to convert these justifications into testable hypotheses. We adopt a model of the type typically used to test for the effects of controls programs: a price equation based on a markup over costs and a wage equation that is consistent with but does not necessarily imply the natural-rate hypothesis.

\*Chase Econometrics, 150 Monument Rd., Bala Cynwyd, PA 19004, and Department of Economics, New York University, New York, NY 10003, respectively. We thank Michael Holdowsky, an anonymous referee, and especially Rob Engle for helpful comments, and Chase Econometrics and C. V. Starr Center for Applied Economics for technical support.

<sup>1</sup>The robustness of the recent new-neoclassical models of inflation is a currently contentious issue, but these models have not yet been employed to test for the effect of controls.

While the focus of our analysis is the Carter program,<sup>2</sup> we also provide additional tests of the effectiveness of the guideposts and the ESP. Our results modify slightly previous conclusions about the ESP: in our view, the catch-up occurred not after the controls were lifted (as Robert Gordon, 1975, and Alan Blinder and William Newton 1981, conclude) but, rather, during the period of decontrol (Phases III and IV). We also show that correction of a specification error in the wage equations typically used in the analysis of controls programs substantially reduces previous estimates (for example, George Perry, 1970, 1980) of the effect and statistical significance of the Kennedy-Johnson guideposts program.

Our principal conclusion about the Carter program is that the roughly concurrent 1979–80 energy-price explosion was not passed through to wages, and hence the general price level, in the usual manner. Whether the moderate behavior of labor costs during this era of double-digit inflation was the result of the Carter program remains an open question.<sup>3</sup>

Section I formulates hypotheses about the effects of wage-price controls in the context of a simplified model of the wage-price process. Section II presents our empirical results.

## I. Formulation of Hypothesis Tests

### A. A Simplified Wage-Price Model

A simplified two-equation model of the wage-price process is as follows:

$$(1) \quad \dot{W} = a_0 + a_1 L(\dot{P}) + a_2 (U - \bar{U});$$

<sup>2</sup>For an extensive description and evaluation of this program, see the final report of the Council on Wage and Price Stability (1981). (We should alert readers that we participated in the administration of this program and the writing of the final report.)

<sup>3</sup>A finding of effectiveness in retarding the inflation rate is not, of course, an endorsement of an incomes policy: against these benefits, one would have to weigh the (unfortunately nonquantifiable) costs of administrative burden and market distortion.

$$(2) \quad \dot{P} = b_0 + b_1 (\dot{W} - R) + b_2 \dot{M}.$$

The wage equation expresses the percentage rate of change of hourly labor compensation,  $\dot{W}$ , as a linear function of a distributed lag on past percentage rates of change of consumer prices,  $L(\dot{P})$ , and the disparity between the actual and natural unemployment rates,  $U - \bar{U}$ .

The distributed lag on past inflation rates is generally interpreted as a reflection of adaptive expectations of future inflation rates. In this interpretation, the coefficient  $a_1$  would be equal to 1.0 in the absence of money illusion (the weights in  $L(\dot{P})$  are normalized to sum to one).<sup>4</sup> We prefer to interpret the inclusion of past inflation rates in the wage equation as reflecting the backward-looking efforts of workers to make up for past inflation, not only to avoid getting sidetracked into the raging debate on the appropriate formulation of an expectations hypothesis, but more fundamentally because we believe that this is a more realistic representation of the formation of wage demands and of compensation practices. Formal cost-of-living adjustment clauses are explicitly backward looking, with today's adjustments depending on past inflation rates. In addition, formal annual pay plans in the non-union sector are typically calibrated to the past year's increase in the cost of living. In testing for effects on expectations, however, we interpret  $L(\dot{P})$  as adaptive expectations. In any event, the above specification of the wage equation cannot discern between forward-looking and backward-looking hypotheses about the determination of wage changes.

The second explanatory variable in equation (1) is a measure of labor-market disequilibrium. Measuring labor-market conditions by the disparity between the actual and natural unemployment rates incorporates into the

<sup>4</sup>Of course,  $a_1 = 0$  implies money illusion only if expected inflation is identically equal to  $L(\dot{P})$ —i.e., that expectations are “extrapolative” as in Philip Cagan (1956) rather than “regressive” (predicated on a “normal” price level), as in Robert Lucas and Leonard Rapping (1969, p. 732).

model the intertemporal shift of the short-run Phillips curve attributable primarily to demographic changes in the labor force (which, until recently, have raised the natural unemployment rate). The coefficient,  $a_2$ , is the slope of the short-run (one-period) Phillips curve—the short-run tradeoff between unemployment and wage inflation (abstracting from one-period feedbacks between the wage and price equations).

So long as  $a_1 = 1$ , the constant term,  $a_0$ , can be interpreted as equilibrium wage growth—that is, the steady-state growth of real wages with equality of the natural and actual unemployment rates. (If we let  $\dot{P}$  equal a constant,  $a_1 = 1$ , and  $U = \bar{U}$ , equation (1) becomes  $\dot{W} - \dot{P} = a_0$ .) In a regime of constant factor shares, the equilibrium wage growth is equal to the trend growth of labor productivity.

A fully specified wage equation includes a number of other variables, including changes in the minimum wage rate and employment taxes.

The price equation (2) is a fairly standard mark-up formula. The percentage rate of increase of prices is a function of the growth of unit labor cost at trend productivity growth,  $R$ , and the percentage rate of change of exogenous materials prices,  $\dot{M}$ . In this specification, capital costs are omitted; hence, the estimated values of  $b_1$  and  $b_2$  should approximately equal the shares of labor and materials in total cost divided by the complement of the capital share. In a regime of approximately constant secular factor shares, the constant term should be close to zero.

One can include capital costs (in which case the equation comes close to being an accounting identity) and a measure of disequilibrium in product markets, but these variables typically have little explanatory power (presumably because of difficulties of measuring capital costs and because inventory fluctuations provide a buffer between market disequilibrium and prices). A variable that typically does have explanatory power is the deviation between actual and trend productivity growth, reflecting the effect of current as well as trend changes in unit labor costs.

## B. Hypothesis Tests

Wage-price models of the above type have been used by a number of investigators to test the effectiveness of incomes policies in slowing the rate of inflation. The usual approach attempts to incorporate the program directly into the model by including a variable that represents it. Typically this is an intercept dummy variable, equal to zero when the program is not in effect and equal to one when it is. The dummy variable can enter the wage equation, the price equation, or some quasi-reduced form (typically with a lagged dependent variable in order to circumvent the problem caused by the infinite lag in the true reduced form). As William Nordhaus (1975) has complained, this approach is extraordinarily unimaginative.

Some investigators have improved upon the straight dummy-variable approach by using multiple dummy variables to reflect different stages of a program or a postcontrols catch-up. The most sophisticated attempt to model controls is that of Blinder and Newton, who construct a variable representing the share of the (nonfood, nonenergy) Consumer Price Index (*CPI*) covered by the ESP during each month of the program. This is an improvement on the typical approach, but it still presumes that the program affects the price equation only through shifts in the intercept term (though by varying amounts each month).

More important, it is questionable whether a coverage variable adequately captures the characteristics of the program. A much more important consideration than coverage is the tightness of the standards and the degree of enforcement. More significant than coverage in the ESP is the distinction between the two freezes, the period of mandatory but more flexible controls (Phase II), and the periods of "voluntary" controls (Phases III and IV). Similarly, in the Carter program, coverage was essentially unvarying, but there was nevertheless a big difference between the first and second years of the program. During the second year, the newly established Pay Advisory Committee substantially relaxed the wage standard and also blocked

enforcement efforts by the Council on Wage and Price Stability. Finally, Perry (1970, 1980), in the most thorough econometric analysis of the Kennedy-Johnson guideposts, phases a dummy variable in and out over the 1962–67 period to reflect varying degrees of intensity despite the fact that coverage was invariant. Thus, coverage typically is not a good indicator of the effectiveness of the design and institutional characteristics of a program. Appropriately constructed dummy variables can usually capture the typically discrete changes in these characteristics.<sup>5</sup>

An incomes policy could conceivably affect the structure of the wage-price process in a number of ways. Some or all of the coefficients of the explanatory variables in either the wage equation or the price equation might change. Moreover, coefficient changes may vary over time. Since the number of possible hypothesis tests is limitless, some thought must be put into determining which hypotheses are worthy of investigation.

1. *Bringing Wage Demands into Line with Productivity Growth.* The typical intercept-shift hypothesis test is justifiable because one commonly stated objective of incomes policies is to educate the public about the relationships between productivity growth, inflation, and income shares. This was the avowed purpose of the Kennedy-Johnson guidepost program, where wage demands were to be brought into line with the trend rate of growth of labor productivity. If  $a_0$  in the wage equation exceeds trend productivity growth, there is an inflationary bias built into the wage equation.

Thus, an intercept dummy can be used to test the hypothesis that an incomes policy depressed the “productivity factor” in wage settlements.<sup>6</sup> Unfortunately, the intercept

dummy tests that have been employed by most investigators are not appropriate because of misspecification of the wage equation, since the trend rate of growth of productivity has been falling over time. If the intercept dummy and the missing productivity variable are correlated, the estimated coefficient of the dummy variable is biased.

2. *Changing the Inflation-Unemployment Tradeoff.* Much of the rhetoric surrounding incomes policies focuses not on shifting the Phillips curve—the effect of an intercept shift—but rather on changing its slope. This was especially notable in the early justification by the Carter Administration, where much of the emphasis was on “...creating an environment in which the effect of fiscal and monetary restraint on unemployment and real growth will be minimized and the effect on inflation will be maximized” (Council on Wage and Price Stability, 1979, p. ix). In this characterization, the incomes policy is intended to operate on the coefficient  $a_2$  in equations (1) and (3). One can test this hypothesis by the incorporation of an interactive variable for the incomes policy,  $I$ :

$$(3) \quad \dot{W} = a_0 + a_1 L(\dot{P}) + (a_2 + \hat{a}_2 I)(U - \bar{U}).$$

This approach, however, is simplistic; presumably, while the objective of an incomes policy may be to steepen the slope of the short-run Phillips curve for *increases* in the unemployment rate, steepening the slope for *decreases* in the unemployment rate would be perverse. The latter effect would imply that the policy exacerbates the inflationary effects of stimulative fiscal or monetary policies. Thus, one might wish to test for an asymmetric effect on the slope of the short-run Phillips curve. For example, the program might increase its slope for higher unemployment rates but leave the slope unchanged for lower unemployment rates. Alternatively, one might hypothesize that the incomes policy would improve the tradeoff in both directions, mitigating the inflationary effects of expansionary policy.

Tests for such asymmetric effects can be formulated, but they require observations during incomes-policy periods with both ex-

<sup>5</sup>As Jon Frye and Gordon (1981) show, a pair of dummy variables does as good a job of reflecting the effects of ESP as the Blinder-Newton coverage variable.

<sup>6</sup>Similarly, an effort by employers to expand profit margins (thus increasing capital's share of total income) would be reflected by a positive value of the constant term,  $b_0$ , in the price equation. Thus, a programmatic intercept dummy in the price equation can be used to test the hypothesis that a price guideline brought profit margins into line with labor and materials-cost growth.

pansionary and contractionary fiscal and monetary policies. Unfortunately, during all three of the incomes policies adopted in the United States in the last two decades, fiscal and monetary policies were primarily expansionary; hence, it is probably not possible to test for a change in the slope of the short-run Phillips curve for higher levels of the unemployment rates.

3. *Deflating Inflationary Expectations.* Perhaps the most common justification for incomes policies (or at least the most respectable among professional economists) is the attempt to deflate inflationary expectations.<sup>7</sup> Interpreting  $a_1 L(\dot{P})$  in equation (1) as an expectations-formation mechanism, we can test for this phenomenon by estimating the effect of a program on  $a_1$  or the parameters of  $L(\dot{P})$ . The latter changes the pattern of adaptation to past inflation rates, whereas the former shifts the lag parameters proportionately.

4. *Insulating the Economy from Shocks.* In 1979, as monetary accommodation of the worldwide oil-price shock allowed domestic inflation rates to hit record peacetime highs, the avowed objective of the Carter Administration's program changed. The original goal of price deceleration was abandoned and replaced by the goal of preventing the energy-price explosion from getting built into the wage-price spiral. Holding firmly to the 7 percent pay guideline, the Carter Administration repeatedly stressed the temporary nature of the shock and the futility of trying to recover, by the escalation of wages, the real income transferred to oil-exporting countries. Quantitative estimates of the direct contribution of the energy-price explosion to the inflation rate were repeatedly underscored.

Testing for the success of an incomes policy in insulating an economy from international price shocks is less straightforward than the above tests for parameter shifts. The test is

not for an effect on the structure of the wage-price model, but for an effect on agents' perceptions of the variables. One possible formulation of this test in the wage equation is

$$\dot{W} = a_0 + a_1 \sum_i \lambda_i \dot{P}_{-i} + \alpha \sum_i \lambda_i S_{-i} + a_2 (U - \bar{U}),$$

where the  $\lambda_i$  are the coefficients of the lag distribution,  $L(\dot{P})$ , the subscript  $-i$  denotes an  $i$ -period lag,  $\alpha$  is a parameter, and  $S$  is the shock component. In particular, the shock component would be equal to zero in all periods except the period of the shock (an admittedly subjective determination), and it would be equal to the contribution to the inflation rate during the shock period (see Section II for a specific application). If the shock component of inflation were not passed through to wage inflation in the usual manner, then  $\alpha$  would be negative. Of course, the maintained hypothesis is that any non-pass-through is attributable to the incomes policy. At a minimum, this requires rough concurrence of the shock and the program.

## II. Empirical Results

Using quarterly U.S. data from the second quarter of 1954 (1954:II) to the second quarter of 1983 (1983:II), we test several of the foregoing hypotheses regarding the effectiveness of the Carter Administration's Pay and Price Standard Program. Precise definitions of the variables used and citations of the sources of data are in the Appendix.

### A. Wage Equations

Table 1 contains the parameter estimates of seven different wage equations. In each, the dependent variable is the percentage change in total labor compensation per hour.

Equation (1.1) incorporates intercept shifts for each of the three controls programs in effect during the sample period. The estimated constant term of 3.2 is an estimate of the equilibrium rate of wage growth. This estimate is precisely equal to the Kennedy-

<sup>7</sup>The Council on Wage and Price Stability stressed the attempt to "...moderate the inflationary expectations of business and worker..." (1979, p. ix).

TABLE 1—WAGE EQUATIONS<sup>a</sup>

	(1.1)	(1.2)	(1.3)	(1.4)	(1.5)	(1.6)	(1.7)
Constant	3.22 (17.65)	3.22 (17.68)					
$U - \bar{U}$	-0.55 (-6.76)	-0.55 (-6.78)	-0.60 (-7.06)	-0.60 (-7.06)	-0.61 (-7.01)	-0.60 (-7.04)	-0.62 (-7.63)
$L(\dot{P})^b$	0.64 (17.86)	0.65 (18.0)	0.84 (28.07)	0.84 (28.07)	0.82 (29.14)	0.84 (28.20)	0.90 (13.42)
DATADUM	-1.41 (-1.54)	-1.40 (-1.54)	-1.38 (-1.44)	-1.38 (-1.44)	-1.40 (-1.42)	-1.38 (-1.44)	-1.40 (-1.55)
CHSSTAX	1.01 (8.21)	1.02 (8.32)	1.01 (7.78)	1.01 (7.78)	1.00 (7.50)	(1.01) (7.78)	0.99 (8.14)
CHMINWAGE	0.02 (3.95)	0.02 (3.97)	0.02 (4.05)	0.02 (4.05)	0.02 (4.07)	0.02 (4.00)	0.02 (3.83)
GUIDEPOSTS	-0.99 (-3.81)	-1.00 (-3.86)	-0.61 (-2.30)	-0.61 (-2.30)	-0.61 (-2.31)	-0.61 (-2.31)	-0.50 (-2.01)
ESPI	0.62 (1.49)	0.61 (1.48)	0.86 (1.97)	0.86 (1.97)	0.86 (1.93)	0.86 (1.98)	0.46 (0.99)
ESP2	-0.77 (-1.59)	-0.85 (-1.77)	-0.61 (-1.19)	-0.61 (-1.19)	-0.31 (-0.61)	-0.63 (-1.23)	-0.48 (-1.06)
STANDARDS	-0.90 (-2.26)						
STANDARDS1		-1.33 (-2.68)	-1.23 (-2.33)				
STANDARDS2		-0.82 (-1.40)	-1.00 (-1.63)				
R			-1.19 (16.41)	1.19 (16.41)	1.22 (16.59)	1.19 (16.44)	1.14 (16.08)
STANDARDS1 · R				-0.93 (-2.32)			
STANDARDS2 · R				-0.80 (-1.64)			
STANDARDS1 · (U - $\bar{U}$ )					4.95 (1.11)		
STANDARDS2 · (U - $\bar{U}$ )					-0.28 (-0.63)		
STANDARDS1 · L( $\dot{P}$ ) <sup>b</sup>						-0.17 (-2.56)	
STANDARDS2 · L( $\dot{P}$ ) <sup>b</sup>						-0.11 (-1.74)	
L(ENERGY - $\dot{P}$ ) <sup>b</sup>							0.01 (0.22)
EXPLOSION							-0.18 (-2.76)
Corrected R <sup>2</sup>	0.862	0.864	0.980	0.980	0.979	0.980	0.982
Standard Error	0.885	0.879	0.929	0.929	0.950	0.927	0.871
D-W Statistic	1.65	1.65	1.56	1.56	1.46	1.57	1.68

<sup>a</sup> The dependent variable is  $\dot{W}$ ; the sample period is 1954:II–1983:III; *t*-statistics are shown in parentheses.

<sup>b</sup>  $L(\cdot)$  connotes a third-degree polynomial distributed lag, with the far endpoint constrained to zero. The lag length is 12 quarters, starting with the variable lagged 1 quarter.

Johnson wage guidepost, an explicit estimate of equilibrium (noninflationary) wage growth at that time.

The estimated coefficient of the disparity between the actual and natural unemployment rates,  $U - \bar{U}$ , is highly statistically significant and is consistent with other esti-

mates of the short-run wage-unemployment rate tradeoff (see, for example, Gordon, 1981; Arthur Okun, 1975; and Perry, 1980). It implies that a one-percentage-point increase in the unemployment rate lowers the rate of wage inflation by half a percentage point. In other words, to lower the rate of wage infla-



tion by one percentage point in a single quarter, the unemployment rate would have to rise by two percentage points. Of course, a sustained rise in the unemployment rate would have an accumulating effect on the wage inflation rate.

The third coefficient is the sum of the coefficients of the distributed lag on prices. The estimate, 0.64, together with the *t*-statistic, implies statistically significant money illusion and a downward-sloping long-run Phillips curve.

The *DATADUM* variable is simply a dummy variable that corrects for the effect of linking two different wage series.

The fifth explanatory variable reflects the effect of changes in employment taxes. The coefficient implies that, at least in the short run, all of the employer component of a Social Security tax increase is shifted forward (higher prices) and none is shifted backward (lower wages).

The sixth variable is the percentage change in the minimum wage. Although statistically significant, the coefficient is extraordinarily small, implying that a ten-percentage-point change in the minimum wage raises the overall wage level (including fringe benefits) by only two-tenths of one percentage point.

The seventh explanatory variable is an intercept-shift dummy to capture the effect of the Kennedy-Johnson guidepost program. We adopt the approach of Perry (1970, 1980), phasing the guideposts in during 1962 and phasing them out during 1967. The coefficient, which is consistent with Perry's estimate, implies that the guidepost program lowered the rate of wage inflation by one percentage point. Moreover, the estimated effect of the program is statistically significant.

The next two variables, *ESP1* and *ESP2*, are intercept-shift dummies for Phases I and II and Phases III and IV, respectively, of the Nixon Administration's ESP. (Phases I and II were periods of strict mandatory control, whereas Phases III and IV were periods of decontrol.) The two dummy variables divide the Nixon controls program approximately in half; consequently, the two coefficients can be added to obtain a rough estimate of the net effect of the program.

The estimates suggest that the Nixon controls program directly raised wage inflation slightly during the strict control phases and depressed wage inflation during the period of decontrol. This apparent paradox can be explained as follows: the perverse *ESP1* estimate reflects the fact that the lowered price inflation in Phases I and II (documented below) was not translated into a commensurately lower rate of wage inflation, and the negative coefficient on *ESP2* reflects the fact that the *ESP2* price surge (documented below) was not passed through in the form of more rapid wage inflation. In other words, on balance, the program had no effect on wages. In any event, the coefficients on *ESP1* and *ESP2* are not statistically significant.

Finally, the *STANDARDS* coefficient and the associated *t*-statistic are consistent with our impressions formed during our involvement in that program. Specifically, our own impression, based in part on extensive discussions with compensation experts and others in the private sector, was that the program lowered the rate of wage inflation by about one percentage point (the coefficient is  $-0.9$ ); but we did not hold this view with a great deal of confidence (the *t*-statistic is  $-2.3$ ).

The Carter Administration's program passed through two distinctive phases: the year of voluntary yet formal controls that were taken quite seriously by the larger U.S. corporations, and then a year of phased decontrol in the context of deteriorating support for the program. To test the hypothesis that the two phases of the Carter program had dissimilar effects, we introduce in equation (1.2) different intercept dummies for the two phases. Our impression is confirmed both by the larger coefficient and the larger *t*-statistic in the first year of the Carter program, and by commensurately lower estimates for the second year. In fact, these estimates pretty much confirm our view that the program lost its *direct* effectiveness in the second year, although lingering feedback effects from the effectiveness of the first year's program did retard wage inflation in the second year. In the tests that follow, we retain the distinction between these two phases of the Carter program, but similar results are obtained by

leaving out *STANDARDS2* or by combining the two phases as in equation (1.1).

It is our view that the standard wage equations, (1.1) and (1.2), are misspecified. Specifically, as noted above, in a wage-price model with no inflationary bias, the constant term should be equal to the trend rate of growth of labor productivity. The problem is that trend productivity growth has not been constant in the United States. Labor productivity in the private business sector was growing at the rate of about 3 percent in the late 1950's, but fell to 2 percent by the end of the 1960's and to 1 percent by the end of the 1970's.

To correct this misspecification, we replace the constant term with a trend-productivity variable (constructed in the manner of Gordon, 1979) in equation (1.3). The coefficient is close to, but statistically significantly larger than, 1.0, the expected value in an equation without inflationary bias.<sup>8</sup>

The correction of the common misspecification of the wage equation blurs somewhat the distinction between the two years of the Carter program, but the most dramatic changes are in other estimated coefficients of the wage equation. First, the sum of the coefficients of lagged prices is increased from 0.6 in equations (1.1) and (1.2) to 0.8 in equation (1.3), and its statistical significance is enhanced. Apparently, the substantial money illusion implied by equation (1.1) and (1.2) is partly attributable to specification error (inflation and trend productivity growth are negatively correlated over the sample period). Another interesting implication of the correction of the specification error in equations (1.1) and (1.2) is that the estimated

effect of the Kennedy-Johnson guidepost program is reduced by 40 percent and its statistical significance is lowered substantially.

When the specification error in equations (1.1) and (1.2) is corrected, the appropriate test for the effect of an incomes program in lowering the productivity factor in wage equations is a dummy variable that interacts with the trend-productivity variable. In equation (1.4) the Carter Administration Pay and Price Standards Program is incorporated with two interactive dummies: *STANDARDS1* · *R* and *STANDARDS2* · *R*. The estimated effect of the program in this equation is -1.2 percentage points in the first year (the coefficient, -.93, multiplied by the average value of *R* during the year, 1.32 percent) and -1.0 percent in the second year (-0.8 × 1.26)—virtually identical to the equation (1.3) estimates.

Equation (1.5) tests the hypothesis that the Carter program changed the slope of the short-run Phillips curve by incorporating interactive dummy variables, *STANDARDS1* · (*U* -  $\bar{U}$ ) and *STANDARDS2* · (*U* -  $\bar{U}$ ). As neither coefficient is statistically significant, we eschew interpretation. (During the *STANDARDS1* period, there was virtually no variation in (*U* -  $\bar{U}$ ); during *STANDARDS2*, (*U* -  $\bar{U}$ ) jumped from 0.0 to 1.8.)

In equation (1.6), we test the hypothesis that the standards program retarded inflationary expectations, under the assumption that *L*( $\dot{P}$ ) is an indicator of workers' expectations of future inflation. The approach taken is to shift the entire lag distribution by the same proportional factor through the incorporation of interactive dummy variables.<sup>9</sup> For the two phases of the program,

<sup>8</sup>We use the uncentered moments matrix in calculating coefficient estimates and variations of the dependent variable about zero (rather than the mean) in calculating the corrected  $R^2$ s in equations (1.3)–(2.12) of Tables 1 and 2. This results in (relatively) efficient coefficient estimates (since a priori information about the constant term is taken into account) and in corrected  $R^2$ s that necessarily fall between 0 and 1. The corrected  $R^2$ s of equations (1.3)–(2.12) are not, however, comparable to those for equations (1.1) and (1.2). (See, for example, Phoebus Dhyrnes, 1978, pp. 21–24.)

<sup>9</sup>The parameters of the distributed lag and the interactive dummy variables were obtained by iterating between the two. The convergence criterion was  $|\hat{\theta}_j^{\tau} - \hat{\theta}_j^{\tau-1}| < .0001$  for all *j*, where  $\hat{\theta}_j^{\tau}$  is the estimate of the *j*th coefficient from iteration  $\tau$ . To obtain a consistent estimate of the variance-covariance matrix of the estimated parameters, we used the method of artificial regressions suggested recently by Russel Davidson and James MacKinnon (forthcoming). In particular, we regressed the residuals from equation (1.6) on the deriva-

the maintained assumption is that the pattern of the lag distribution is unaffected. The estimated coefficients of equation (1.6) imply that the standards program lowered inflation expectations by 17 percent in the first year and 11 percent in the second. As  $L(\dot{P})$  over the relevant periods averaged 7.4 percent and 9.7 percent, this estimate translates to direct effects on wage inflation of 1.3 and 1.1 percentage points during the two phases.

The final hypothesis test regarding the effectiveness of the Carter controls program is a test for the Carter Administration's success in preventing the energy price explosion of 1979-80 from getting built into wage demands. In particular, we test the hypothesis that the workers did not take into account price increases directly attributable to the energy-price explosion in formulating their wage demands.

Our approach is to estimate the coefficient of

#### EXPLOSION

$$= \sum_{i=1}^{12} \lambda_i (ENERGY - \dot{P})_{-i} \delta_{-i},$$

where  $\lambda_i, i=1, \dots, 12$ , are the estimated coefficients of the distributed lag on prices,  $L(\dot{P})$ ;  $(ENERGY - \dot{P})_{-i}$  is the difference between the increases in energy prices and the *CPI*, lagged  $i$  quarters; and  $\delta$  is a dummy variable equal to 1 from 1979:I to 1980:II

and zero otherwise.<sup>10</sup> If the energy price increases of 1979-80 were treated exactly as other price increases, the coefficient of *EXPLOSION* would be zero; if these increases in excess of the rise of other components of the *CPI* were entirely swallowed by workers, the coefficient would be equal to  $\omega/(1-\omega)$ , where  $\omega$  is the relative importance of energy prices in the *CPI* ( $\omega/(1-\omega) = .115$  in 1979-80).<sup>11</sup>

To test the hypothesis that the Carter program prevented a full pass-through of the energy price explosion, we must confirm that workers typically treat energy price increases just as they treat any other price increase in the formulation of wage demands. We do this by the inclusion in equation (1.7) of a distributed lag on the difference between changes in energy prices and the overall inflation rate, denoted  $L(ENERGY - \dot{P})$ . If workers typically treat energy price increases no differently than increases in other components of the *CPI*, the coefficient of this difference would be zero. As can be seen from Table 1, the estimated coefficient is statistically insignificant.

The coefficient on *EXPLOSION* is remarkably large and statistically significant. The coefficient is more than  $1\frac{1}{2}$  times  $\omega(1-\omega)$ , which is consistent with the hypothesis that much of the indirect effect of the energy price explosion, as well as the direct effect, was not passed through in the form of higher wage demands. Equation (1.7) has the lowest standard error in Table 1. It is also superior to the other equations in two other im-

portant respects of equation (1.6) with respect to the parameters, evaluated at their estimated values. The estimated coefficients of this artificial regression were identically equal to zero (up to rounding error), indicating that the iterative process converged to a maximum of the likelihood function. The variance-covariance matrix of this artificial regression provides a consistent estimate of the variance-covariance matrix of the original regression, equation (1.6), and is used to calculate the  $t$ -statistics reported in Table 1. The Davidson-MacKinnon correction made little difference in equation (1.6); most of the  $t$ -statistics were virtually unaffected, the only perceptible changes being small (absolute) increases in the  $t$ -statistics for *STANDARDS1·R* (from -2.43 to -2.56) and *STANDARDS2·R* (from -1.68 to -1.74). In other wage equations discussed below, however, the correction resulted in substantial changes in many estimates—especially the standard errors of particular interest.

<sup>10</sup>Percentage increases in the energy component of the *CPI* during 1978, 1979, and 1980 were as follows:

	I	II	III	IV
1978	3.6	9.7	10.9	5.8
1979	16.8	52.5	62.4	19.9
1980	46.6	26.5	8.1	-2.4

<sup>11</sup>Estimation of the coefficient of *EXPLOSION* required an iteration between it and the coefficients of  $L(\dot{P})$ . The approach taken was identical to that used in estimating equation (1.6). The Davidson-MacKinnon correction lowered the  $t$ -statistic on  $L(\dot{P})$  from 25.69 to 13.42 and that on *EXPLOSION* from -4.35 to -2.76.

TABLE 2—WAGE-EQUATION SENSITIVITY ANALYSIS<sup>a</sup>

	(1.7)	(2.8)	(2.9)	(2.10)	(2.11)	(2.12)
<i>R</i>	1.14 (16.08)	-0.96 (-1.34)	1.14 (16.12)	1.86 (6.58)	1.63 (5.60)	1.09 (12.12)
$U - \bar{U}$	-0.62 (-7.63)	-0.57 (-7.15)	-0.63 (-7.59)	-0.49 (-4.09)	-0.52 (-4.71)	-0.58 (-7.23)
$L(\dot{P})^b$	0.90 (13.42)	0.50 (3.49)	0.90 (18.14)		0.80 (9.75)	0.91 (6.13)
<i>DATADUM</i>	-1.40 (-1.55)	-1.40 (-1.60)	-1.52 (-1.68)	-1.44 (-1.73)	-1.43 (-1.73)	-1.40 (-1.65)
<i>CHSSTAX</i>	0.99 (8.14)	1.03 (8.64)	0.97 (7.83)	1.11 (8.34)	1.07 (8.22)	1.00 (8.44)
<i>CHMINWAGE</i>	0.02 (3.83)	0.01 (3.10)	0.02 (3.78)	0.01 (1.75)	0.01 (2.35)	0.02 (3.65)
<i>GUIDEPOSTS</i>	-0.50 (-2.01)	-1.15 (-3.61)	-0.54 (-2.15)	-1.66 (-3.35)	-1.32 (-2.62)	-0.38 (-1.47)
<i>ESP1</i>	0.46 (0.99)	0.27 (0.60)	0.41 (0.89)	-0.12 (-0.28)	-0.21 (-0.47)	-0.29 (-0.56)
<i>ESP2</i>	-0.48 (-1.06)	-0.85 (-1.87)	-0.48 (-1.05)	-1.13 (-2.45)	-0.62 (-1.38)	-0.40 (-0.80)
$L(ENERGY - \dot{P})^b$	0.01 (0.22)	0.07 (1.14)	0.02 (0.27)		0.02 (0.28)	-0.02 (-0.25)
<i>EXPLOSION</i>	-0.18 (-2.76)	-0.27 (-2.13)	-0.18 (-2.74)	-0.08 (-0.83)	-0.17 (-2.03)	-0.21 (-2.32)
Constant		5.73 (2.95)				
<i>PRODDEV</i>			0.03 (1.11)			
$L(HOMEOWN - \dot{P})^b$						0.08 (0.78)
$L(CHPCE)^b$				0.81 (7.12)		
$L(ENERGY - CHPCE)^b$				0.02 (0.23)		
Corrected $R^2$	0.982	0.982	0.982	0.988	0.988	0.985
Standard Error	0.871	0.849	0.870	0.806	0.799	0.825
<i>D-W</i> Statistic	1.68	1.66	1.69	1.99	2.11	1.87

<sup>a</sup>The dependent variable is  $\dot{W}$ ; the sample period is 1954:II–1983:II for equations (2.7)–(2.9), 1962:II–1983:II for equations (2.10) and (2.11), and 1956:II–1983:II for equation (2.12); *t*-statistics are shown in parentheses.

<sup>b</sup>See Table 1.

portant respects. First, the coefficient of trend productivity growth  $R$  is closer to 1.0 and, in fact, is (barely) statistically insignificantly different from 1.0. Second, the sum of the coefficients on lagged prices (0.9) is higher (though still statistically significantly less than 1.0).

### B. Wage-Equation Sensitivity Analysis

Because wage equations are notoriously nonrobust, we have estimated several alternative wage equations to test for sensitivity of the strong result in equation (1.7) to changes in its specification.

For easy comparison, equation (1.7) is reproduced in Table 2.<sup>12</sup> The first sensitivity test, shown in equation (2.8), indicates that a constant term added to the basic equation (1.7) is statistically significant. Moreover, addition of the constant term makes the coefficient of the trend-productivity variable,  $R$ , statistically insignificant. Equation (2.8), however, is much less plausible, primarily be-

<sup>12</sup>The *t*-statistics in Table 2 are all calculated using the artificial-regression method described above. In every case, the Davidson-MacKinnon correction cut the *t*-statistic on *EXPLOSION* roughly in half.

cause of the resurrection of substantial money illusion, as indicated by the decline in the sum of the coefficients on lagged prices from .9 to .5. In addition, the *t*-statistic on the sum of the price coefficients falls from 13.4 to 3.5. Theory tells us that the constant term does not belong in this equation.

Equation (2.9) tests for the possible effect of cyclical, in addition to trend, productivity movements on rates of wage increase by introducing the difference between actual and trend productivity as an explanatory variable (*PRODDEV*). The coefficient is insignificantly different from zero, indicating that wage growth is not affected by cyclical productivity changes.

Potentially the most troubling sensitivity test that we carried out was the replacement of the *CPI* by the Fixed Weighted Personal Consumption Expenditure (*PCE*) Price Index from the National Income Accounts. The resultant regression, which must be run on a shorter sample period, is reported as equation (2.10) in Table 2. The substitution substantially lowers the point estimate and eliminates the statistical significance of the effect of the Carter program in preventing the energy price explosion from getting built into the wage-price process.

Equation (2.11) is the basic equation (1.7) run on the shorter sample period of equation (2.10). The large and statistically significant estimate of the effect of the Carter program in equation (2.11) indicates that the diminution in the estimated effect in equation (2.10) is attributable to the substitution of the *PCE* Price Index for the *CPI*, rather than the shortening of the sample period.

There is little to choose—either in goodness of fit or plausibility of the coefficient estimates—between equations (2.10) and (2.11).<sup>13</sup> The issue, then, is which price index is appropriate to use in a wage equation—that is, which price index is used by workers in formulating their estimate (or expectation) of changes in the cost of living. It seems

apparent that workers typically use the *CPI* rather than the *PCE* Price Index: the *CPI* is almost always used in formal cost-of-living adjustment clauses, and it is by far the most visible measure of the cost of living. As the principal difference between the two indexes is the treatment of home ownership (the *CPI* until 1983 used home-ownership costs, whereas the *PCE* Price Index uses imputed rent), a test of the hypothesis that workers base wage demands on the *PCE* Price Index rather than the *CPI* is approximately equivalent to a test of a differential response to changes in home-purchase prices and mortgage-interest costs in the *CPI* as compared to changes in other components of *CPI*.

Equation (2.12) tests this hypothesis by including as an explanatory variable a distributed lag on the difference between home-purchase and mortgage-interest costs and the total *CPI*,  $L(\text{HOMEOWN} - \dot{P})$ . If workers treat these costs differently from other costs (as would be the case if they used the *PCE* Price Index in estimating the cost of living), this coefficient would be different from zero. The estimated coefficient and the *t*-statistic decisively reject the hypothesis that workers treat the homeownership component of the *CPI* differently from other components.

On the basis of this test, but more fundamentally on the basis of common sense, we believe that the *CPI* is the appropriate measure of the cost of living to be used in a wage equation. Nevertheless, the fact that replacement of the *CPI* by the *PCE* Price Index lowers substantially the large estimated effect of the Carter program in equation (1.7) is evidence that the nonrobustness of wage equations makes it quite possible to get conflicting results with reasonable alternative specifications.

### C. Price Equation

Various tests of the effect of the Carter program on the structure of the price equation (not reported here) uniformly and decisively reject the hypothesis of effectiveness. Note, however, that to have an effect on inflation, there is no need for a program to affect the structure of the price equation

<sup>13</sup> Both equations are inferior to the equations run on the larger sample period. Especially noteworthy are the large coefficients of *TRENDPROD* in equations (2.10) and (2.11).

TABLE 3—PRICE EQUATION<sup>a</sup>

Constant	-0.06 (-0.28)	<i>WTENERGY</i> (-2)	0.37 (3.35)
$\dot{W} - R$	0.27 (3.87)	<i>WTHOMEOWN</i>	1.05 (9.74)
$\dot{W}(-1) - R(-1)$	0.24 (3.34)	<i>PRODDEV</i>	-0.08 (-2.44)
$\dot{W}(-2) - R(-2)$	0.17 (2.49)	<i>GUIDEPOSTS</i>	0.54 (1.84)
<i>WTENERGY</i>	0.59 (5.38)	<i>ESP1</i>	-0.90 (-1.91)
<i>WTENERGY</i> (-1)	0.05 (0.46)	<i>ESP2</i>	2.47 (4.92)
Corrected $R^2$ 0.933; Standard Error 1.011; $D-W$ Statistic 1.91			

<sup>a</sup>The dependent variable is  $\dot{P}$ ; the sample period is 1954:II–1983:II;  $t$ -statistics are shown in parentheses.

directly. If the structure of the wage equation is changed, a program will lower price inflation indirectly through the pass-through of lower labor-cost inflation in the price equation. Table 3 contains the estimated price equation, to be used in simulations reported below.

The constant term, as expected in the absence of an inflationary bias, is roughly equal to zero.

The percentage change in unit labor costs, evaluated at trend productivity growth, is entered concurrently and with one- and two-period lags. The sum of the coefficients is 0.7, which is about what one would expect in view of the fact that energy prices and home-ownership costs (with relative-importance weights summing to 0.3) are included in the equation as explanatory variables.

The percentage change of energy prices, weighted by its relative importance in the *CPI*, is included both concurrently and with one- and two-period lags (*WTENERGY*). The coefficients sum to 1.0, which suggests that the indirect effects of energy price increases operate through the wage equation.

The variable *WTHOMEOWN* is the (weighted) percentage rate of change of home-purchase and mortgage-interest costs in the *CPI*. It is included not because it is exogenous (for certainly it is not), but rather to correct for inappropriate treatment of the costs of home ownership in the calculation of the *CPI* before January 1983. Most index-number specialists agree that before 1983 the

*CPI* was overly sensitive to changes in mortgage interest rates and home-purchase costs because of conceptual problems with the way that these costs were calculated. The resultant distortion was especially prominent during the last few years before the Bureau of Labor Statistics was forced, by the preponderance of professional opinion, to change its method of calculating home-ownership costs (switching to an imputed-rent approach). Thus, the inclusion of *WTHOMEOWN* can be interpreted as a correction for the earlier mismeasurement of inflation. The coefficient, as expected, is insignificantly different from 1.0.

To test for the possible effect of cyclical productivity changes on inflation, we also include as an explanatory variable the difference between trend and actual productivity growth, *PRODDEV*. The estimated coefficient of this variable implies that 8 percent of the productivity shortfall during a recession is passed through in the form of higher prices. The other 92 percent is absorbed in the fluctuation of profit margins.

The coefficient of the guidepost dummy variable suggests a perverse effect on price inflation. Roughly speaking, the point estimate implies that the labor-cost savings attributable to the guidepost program were not passed through in the form of commensurately lower inflation. The coefficient is not quite statistically significant, however.

The estimated coefficients of the dummy variables, *ESP1* and *ESP2*, indicate that the

Nixon Administration's controls program reduced the rate of price inflation by one percentage point during the tough Phases I and II, but that a price surge during decontrol (Phases III and IV) more than eliminated the earlier salutary effects. The sum of the coefficients of *ESP1* and *ESP2* is positive and statistically significant, implying a perverse effect of the entire program.<sup>14</sup>

Blinder and Newton also found that a postcontrols catch-up eliminated the large initial effects of the Nixon program. Our results, however, indicate that the catch-up came during the decontrol period, rather than after the controls were formally lifted in 1975. This timing contrast is probably attributable to different treatments of energy prices. The postcontrols catch-up is somewhat confounded with the effects of the 1973-74 oil-price explosion. Energy prices took off in the last part of 1973 and surged through the first two quarters of 1974; controls were lifted early in the second quarter of 1974. The dependent variable of Blinder and Newton is the underlying rate of inflation, which excludes energy prices. As a surge in energy prices gets built into the underlying rate with a short lag, we feel that the 1974 price increases attributed by Blinder and Newton to a postcontrols catch-up might in fact be caused by the lagged effect of the energy price explosion.<sup>15</sup>

#### D. Simulations

The direct effect of the Carter program on wage inflation (assuming that the absorption of the energy price increases was attributable to the program rather than to other structural change) can be obtained by simulating the wage equation (1.7). The full effect of the program, taking into account the interaction between prices and wages, is estimated by simulating the wage equation (1.7) and the price equation jointly.

These simulations are depicted in Figures 1-3 and summarized in Table 4. The Pro-

<sup>14</sup> The estimated covariance between the coefficients of *ESP1* and *ESP2* is .003.

<sup>15</sup> A postcontrols dummy is statistically insignificant in our equation.

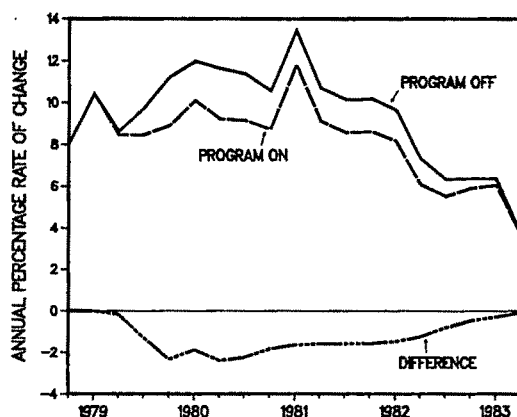


FIGURE 1. SIMULATED DIRECT (PARTIAL) EFFECT ON WAGES

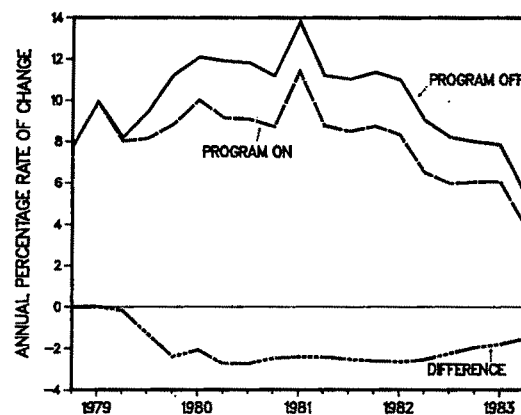


FIGURE 2. SIMULATED FULL EFFECT ON WAGES

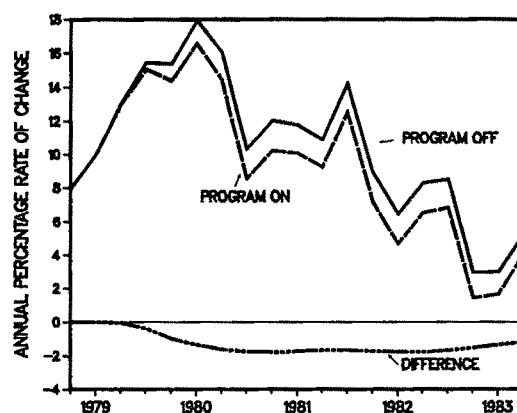


FIGURE 3. SIMULATED FULL EFFECT ON PRICES

TABLE 4—SIMULATIONS<sup>a</sup>

Period	Direct (Partial) Effect on Wages			Full Effect on Wages and Prices					
	Program On	Program Off	Differ- ence	Wages			Prices		
				Program On	Program Off	Differ- ence	Program On	Program Off	Differ- ence
Controls (1978:IV– 1980:IV)	9.06	10.41	–1.35	8.84	10.38	–1.54	12.33	13.12	–0.79
Controls and Postcontrols (1978:IV– 1983:II)	8.16	9.37	–1.21	8.09	10.02	–1.93	9.17	10.44	–1.27

<sup>a</sup>Average annual percentage rates of change.

gram On calculations are predicted percentage changes using equation (1.7) (and, except in Figure 1, the price equation) as estimated; the Program Off calculations are obtained by setting the *EXPLOSION* variable equal to zero in all quarters. The difference between the two percentage changes is the simulated effect of the program. In each case, there is no effect until 1979:II because energy prices did not take off until 1979:I (i.e., the *EXPLOSION* variable is equal to zero before 1979:I) and prices first enter the wage equation with a one-period lag.

The simulated direct effect on wage inflation, illustrated in Figure 1, quickly climbs to two percentage points by the end of 1979 and persists at that level through most of 1980. Through 1981 and 1982, the simulated effect declines slowly but monotonically, virtually vanishing by early 1983. The average direct effect on wages over the course of the Carter program (1978:IV–1980:IV) is 1.3 percentage points. This estimate is roughly consistent with the alternative estimates of the direct effect of the program on wages reflected in equations (1.1)–(1.6) in Table 1.

The more provocative aspect of this simulation is the persistence of the effect until long after the program was formally abandoned in January 1981. This enduring effect is technically attributable to the twelve-quarter lag on prices,<sup>16</sup> and the maintained

hypothesis that workers were persuaded to absorb the energy price increases for all time—that is, no attempt was made to recoup these concessions after the program was abandoned. (An alternative equation that constrains the effect of the program to end in 1980:IV has a significantly worse fit than equation (1.7); various equations incorporating postcontrols catch-up variables instead of a persistent absorption of the energy price increase are also inferior to equation (1.7).)

The joint simulation of the wage equation (1.7) and the price equation indicate that, as would be expected, taking full account of the interaction between wages and prices substantially enhances the simulated effect on wage inflation (Figure 2). Over the period of the Carter program, the simulated effect is increased by only 0.2 percentage points (from 1.3 to 1.5 percentage points), but the big difference is the large simulated effects after

significant all the way to the end of the 12-quarter lag:

Lag	Coefficient	t-Statistic
–1	.16	4.3
–2	.12	7.1
–3	.08	7.5
–4	.07	4.2
–5	.06	3.3
–6	.06	3.6
–7	.06	4.7
–8	.07	6.4
–9	.07	6.4
–10	.07	5.3
–11	.06	4.4
–12	.04	3.8

<sup>16</sup>The mean lag is 4.4 quarters and, although the weights are largest for short lags, coefficients remain



the program was abolished. In fact, the full effect on wages persists at 2 percentage points until mid-1982, and as recently as 1983:II the effect was still  $1\frac{1}{2}$  percentage points.

The simulated effect on price inflation (Figure 3) during the Carter program is 0.9 percentage points, but from early 1980 until early 1983, the simulated effect on the rate of change of the *CPI* is substantially greater than 1 percentage point, sometimes approaching 2 percentage points. (The effect on price inflation is not as large as the effect on wage inflation because the sum of the coefficients on unit labor costs in the price equation is less than 1.0.)

### III. Concluding Remarks

The foregoing tests, like most econometric tests, are unlikely to change the opinions of those with strong priors about the effectiveness of controls. They nevertheless make it clear that, for one reason or another, the energy price surge of 1979–80 was not passed through in the usual manner. As a result, from late 1978 until the middle of 1983, wage inflation was about 2 percentage points lower and price inflation was about 1 percentage point lower than would have been expected on the basis of the structure of our wage-price model.

The foregoing simulations and hypothesis tests maintain that the only change in the structure of the wage-price process in recent years was caused by the Carter Administration Pay and Price Standards Program. It is, of course, possible (some might say likely) that there are other causes of the change in structure. For example, casual inspection indicates that there may have been a worldwide moderation of wage demands in the face of the acceleration of inflation brought about by the 1979–80 energy crisis. Since most countries did not introduce incomes policies during this period, it is possible that a more fundamental structural change took place worldwide and that this accounts for the large estimated effect of the Carter program. On the other hand, a common response to an international shock could reflect the interdependence of Western economies and the transmission of a programmatic effect in the United States alone. A systematic

international comparison would be required to sort out these phenomena.

Alternatively, it is possible that the wage equation is misspecified, in the sense that the effect of price changes on wage demands is nonlinear—in particular, that wage demands increase less than proportionately as inflation grows. There are institutional reasons for this phenomenon—for example, caps on costs-of-living adjustments. It may well be that the reason wages behaved moderately in 1979–80 in the face of rampant inflation brought on by the worldwide energy crisis has nothing to do with the Carter program and is instead attributable to the fact that this inflation was outside the bounds of the postwar U.S. experience and that linear models estimated on the basis of that experience would automatically overpredict wage increases during a period of double-digit inflation. Additional research is required to answer these questions.

### APPENDIX

#### *Variable Definitions and Data Sources*<sup>17</sup>

$\dot{W}$  = percentage change in *PAY*.

$PAY = J \times WSS/WS$ .

*J* = hourly earnings index of production workers in private nonfarm sector. This series is adjusted for overtime (in manufacturing) and for interindustry shifts in employment. 1954–63: Gordon (1971); 1964–83: Bureau of Labor Statistics (BLS).

*WSS* = compensation of employees. National Income Accounts (NIA).

*WS* = wages and salaries. NIA.

$\dot{P}$  = percentage change in the *CPI* for urban consumers (all items). BLS.

*CHPCE* = percentage change in the Fixed Weighted Personal Consumption Expenditures Price Index (all items). Bureau of Economic Analysis.

$U - \bar{U}$  = unemployment rate for civilian workers less the natural unemployment rate calculated by Gordon (1978). BLS.

<sup>17</sup> Unless otherwise noted, all variables are seasonally adjusted. All percentage changes are at annual rates.

$CHSSTAX$  = percentage change in  $1/(1-TWER/WS)$ .

$TWER$  = employer contributions for social insurance. NIA.

$CHMIWAGE$  = percentage change in the minimum hourly wage for all covered and nonexempt workers (not seasonally adjusted). Office of Fair Labor Statistics, Department of Labor.

$GUIDEPOSTS$  = .25 for 1962:I, .5 for 1962:II, .75 for 1962:III, 1 for 1962:IV to 1966:IV, .75 for 1967:I, .5 for 1967:II, .25 for 1967:III, 0 otherwise.

$ESPI$  = .5 for 1971:III, 1 for 1971:IV to 1972:IV, .167 for 1973:I, 0 otherwise.

$ESP2$  = .833 for 1973:I, 1 for 1973:II to 1974:I, .333 for 1974:II, 0 otherwise.

$STANDARDS1$  = .667 for 1978:IV, 1 for 1979:I to 1979:IV, 0 otherwise.

$STANDARDS2$  = 1 for 1980:I to 1980:III, .667 for 1980:IV, 0 otherwise.

$STANDARDS = STANDARDS1 + STANDARDS2$ .

$R$  = estimated trend rate of productivity growth, obtained by regressing  $CHPROD$  on  $CHGAP$ ,  $CHGAP_{-1}$ ,  $TIME$ , and  $DEOE$ , the period of estimation being 1953:I to 1980:III, and using the fitted values setting  $CHGAP$ ,  $CHGAP_{-1}$ , and  $DEOE$  equal to zero.

$CHPROD$  = percentage change in output per man-hour in the nonfarm business section. BLS.

$GNPGAP = ((POTGNP - GNP) / POTGNP) \times 100$ .

$GNP$  = Gross National Product-1972 dollars. NIA.

$POTGNP$  = Potential  $GNP$ -1972 dollars. Council of Economic Advisors.

$CHGAP$  = percentage change in  $GNPGAP$ .

$TIME$  = 1 for 1953:I, 2 for 1953:II, and so on.

$DEOE$  = end-of-expansion dummy constructed by Gordon (1979).

$ENERGY$  = percentage change in the energy component of the  $CPI$  (not seasonally adjusted). Constructed using BLS data.

$WTENERGY = ENERGY$  multiplied by relative importance of energy component of the  $CPI$ .

$WTHOMEOWN$  = weighted percentage change in the home-purchase and mortgage-interest cost component of the  $CPI$ , the weight being its relative importance in the  $CPI$  (not seasonally adjusted). Constructed using BLS data.

$PRODDEV = CHPROD$  less  $TREND-PROD$ .

## REFERENCES

- Blinder, Alan S., and Newton, William J., "The 1971-1974 Controls Program and the Price Level: An Econometric Post-Mortem," *Journal of Monetary Economics*, July 1981, 8, 1-23.
- Cagan, Philip, "The Monetary Dynamics of Hyperinflation," in M. Friedman, ed., *Studies in the Quantity Theory of Money*, Chicago: University of Chicago, 1956.
- Davidson, Russel and MacKinnon, James G., "Model Specification Tests Based on Artificial Linear Regressions," *International Economic Review*, forthcoming.
- Dhrymes, Phoebus J., *Introductory Econometrics*, New York: Springer-Verlag, 1978.
- Frye, Jon F., and Gordon, Robert J., "Government Intervention in the Inflation Process: The Econometrics of 'Self-Inflicted Wounds'," *American Economic Review Proceedings*, May 1981, 71, 288-94.
- Gordon, Robert J., "Inflation in Recession and Recovery," *Brookings Papers on Economic Activity*, 1:1971, Appendix C, 153-58.
- \_\_\_\_\_, "The Impact of Aggregate Demand on Prices," *Brookings Papers on Economic Activity*, 3:1975, 613-62.
- \_\_\_\_\_, *Macroeconomics*, Boston: Little, Brown & Co., 1978.
- \_\_\_\_\_, "The End-of-Expansion Phenomenon in Short-Run Productivity Behavior," *Brookings Papers on Economic Activity*, 2:1979, 447-61.
- \_\_\_\_\_, "Output Fluctuations and Gradual Price Adjustment," *Journal of Economic Literature*, June 1981, 19, 493-530.
- Lucas, Robert E. and Rapping, Leonard A., "Real Wages, Employment, and Inflation," *Journal of Political Economy*, September/October 1969, 77, 721-54.
- Nordhaus, William D., "Comment," *Brookings*

- Papers on Economic Activity* 3:1975, 663-65.
- Okun, Arthur M., "Inflation: Its Mechanics and Welfare Costs," *Brookings Papers on Economic Activity*, 2:1975, 351-90.
- Perry, George L., "Changing Labor Markets and Inflation," *Brookings Papers on Economic Activity*, 3:1970, 411-48.
- \_\_\_\_\_, "Inflation in Theory and Practice," *Brookings Papers on Economic Activity*, 1:1980, 207-41.
- Council on Wage and Price Stability, Executive Office of the President, *Pay and Price Standards: A Compendium*, Washington, USGPO, 1979.
- \_\_\_\_\_, *Evaluation of the Pay and Price Standards Program*, Washington: USGPO, 1981.

# Stardom and Talent

By MOSHE ADLER\*

The phenomenon of stars is defined by Sherwin Rosen to be one "wherein relatively small numbers of people earn enormous amounts of money and dominate the activities in which they engage" (1981, p. 845). Rosen sets out to explain two aspects of this phenomenon: persons with only a slightly greater talent command much higher incomes than those who are only slightly less talented; output is concentrated on those few who have the most talent. Rosen's explanation consists of two factors: lesser talent is a poor substitute for greater talent, and either the activity can be reproduced endlessly (for example, on records) at a fixed cost, or the cost of production does not rise in proportion to the size of the seller's market (a better surgeon can perform better operations and more of them within a given time).

Rosen explains why large differences in earnings could exist where there are only small differences in talent. This paper explains why large differences in earnings could exist even where there are no differences in talent at all. In other words, it explains why there could be stars among individuals known to have equal talents.

## I. Analysis

The main argument of this paper is that the phenomenon of stars exists where consumption requires knowledge. The consumer's utility function is similar in spirit to

the one developed by George Stigler and Gary Becker (1977), where consumers accumulate "consumption capital." As an example, consider listening to music. Appreciation increases with knowledge. But how does one know about music? By listening to it, *and by discussing it with other persons who know about it*. In this learning process lies the key to the phenomenon of stars.

The learning process in this paper adds to the learning process in the Stigler-Becker framework the element of discussion with knowledgeable individuals. By itself, the Stigler-Becker model is sufficient to yield that consumers will not diversify indefinitely either across activities, or across individuals within a given activity. (Stigler and Becker did not make this point.) An opera fan must forego some fields of interest (say, golf, rock music) for the sake of greater knowledge of opera. Once a field of interest is chosen, the fan must forego having very little knowledge about a great number of performers for the sake of greater acquaintance with the work of a few. Each person ends up with a limited number of artistic activities and, within each activity, a limited number of stars. What remains to be explained is why everyone would choose to have the *same* stars.

Here, the need to discuss with other knowledgeable individuals in order to know is essential. If every individual were knowledgeable about a different artist, no discussion would be possible. One is better off patronizing the same artist as others do. It is plausible to assume that the cost of searching for knowledgeable discussants is minimized if one chooses the most popular artist. Thus, if other artists are not cheaper by more than the savings in search costs, one is better off patronizing the star. Alternatively, if other artists are not sufficiently *better*, one is better off patronizing the star. To reemphasize, the star need not possess greater talent. Stardom is a market device to economize on learning costs in activities where "the more you know

\*Assistant Professor, Department of Economics, University of California, Davis, CA 95616. I thank Alanson Minkler for research assistance and substantial criticism and comments. Louis Makowski, Tom Russell and Joaquim Silvestre provided help beyond the call of collegiality. I also thank Sy Adler, Vic Goldberg, Jay Helms, Alan Olmstead, Joe Ostroy, John Roemer, Wendy Sarvasy, Steve Sheffrin, Art Sullivan, Richard Zerbe, the participants in the labor workshop at Stanford and the faculty seminar at Davis and an anonymous referee for helpful comments. This paper was motivated by a dispute with Samira Haj.

the more you enjoy." Thus stardom may be independent of the existence of a hierarchy of talent.

## II. The Definition of Talent<sup>1</sup>

Assume an economy with identical consumers and nonidentical producers called artists, denoted by  $X$ ,  $Y$ , and  $Z$ . Let a consumer's utility from the consumption of art,  $U$ , be independent of the consumption of all other goods.

$$(1) \quad U = U(x, y, z),$$

where  $x$ ,  $y$ , and  $z$  are measured in units of time that the consumer devotes to the art produced by  $X$ ,  $Y$ , and  $Z$ . For music this would be listening time and for paintings, observation time.<sup>2</sup>

Two artists,  $X$  and  $Y$ , are said to have lesser, equal, or greater talent if the utility function satisfies the respective condition:

$$(2) \quad U(x, 0, 0) \leq U(0, y, 0)$$

for all  $x$ ,  $y$  such that  $x = y$ . It is not necessary that the inequal will have the same direction at all levels of  $x = y = z$ . It is assumed, however, that this is the case.

## III. The Model

The simplest model assumes only two artists of equal talent,  $X$  and  $Y$ . Learning about the artists involves direct contact with their work (listening to their music, observing their painting, etc.), and discussing the work with other knowledgeable individuals. Assume that the learning process is of a fixed proportion between direct contact time and discussion time. In my model, there is no distinction between learning and consuming. One enjoys both direct contact with art and discussion of art, or one learns through con-

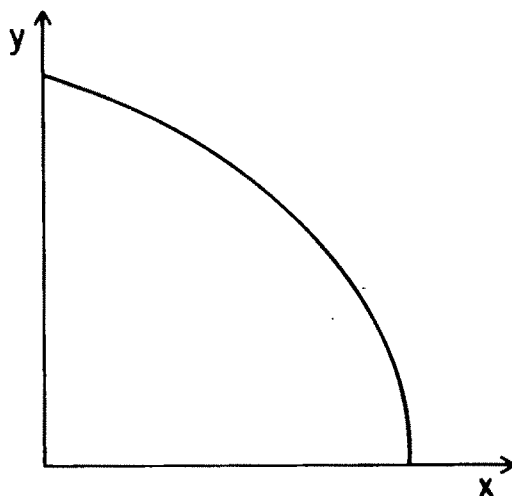


FIGURE 1

sumption. Call the composite good of contact with art and discussion of art—art. The “addictive” nature of art where specialization is preferred yields concave indifference curves. A consumer prefers  $c$  units of  $x$  or  $c$  units of  $y$  to any combination of  $x$  and  $y$  totalling  $c$  units. Figure 1 depicts the indifference map.

To illustrate, the indifference map could be generated by a separable utility function of the form

$$(3) \quad U(x, y) = u(x, 0) + u(0, y).$$

Since  $x$  and  $y$  are of equal talent,  $u$  has only one parameter.

$$(4) \quad u = u(v); \quad v = x, y,$$

if  $x = y$ ,  $u(x) = u(y)$ . Because enjoyment increases with knowledge the marginal utility is increasing:  $u' > 0$ ,  $u'' > 0$ .

To determine the consumer's choice, assume that the only cost in the consumption of art is time. This cost consists of two elements: the actual time devoted to art (direct contact and discussion), and the time devoted to the search for individuals with whom one could discuss the artist one chooses. Assume that the search time is  $1/X$  and  $1/Y$  where capital letters indicate the total number of consumers who choose the

<sup>1</sup>Thanks go to Leon Wegge for insisting that such a definition is required, and to Tom Russell for simplifying my definition.

<sup>2</sup>The time devoted to art includes the time devoted to the discussion of art. See below.

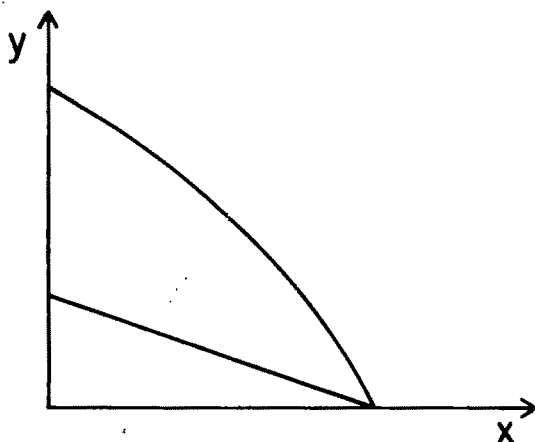


FIGURE 2

corresponding artist. Assume also that this search cost has to be incurred for each unit of art consumed. The consumer devotes  $I$  units of time over his or her life to art and the search involved. The budget constraint of our consumer is

$$(5) \quad x(1 + 1/X) + y(1 + 1/Y) = I.$$

From the budget constraint and the indifference curves it is clear that the consumer will specialize: he or she will either consume  $x$  or  $y$ , not both. If  $X = Y$  the consumer will be indifferent between the two, but if more consumers consume  $x$ ,  $X > Y$ , our consumer will be better off consuming  $x$ . Figure 2 depicts the maximization problem. In this discussion, the existence of a super star among equals is apparent.<sup>3</sup>

Note that whereas all individuals could have equal talents, not all individuals would be artists. An artist for the purpose of this paper is one who produces a good with the

quality of increasing marginal utility in consumption. Only artists could be stars.

#### A. More than One Star

In the world of art, there is more than one star. A minor modification in the model would result in multiple stars. Assume that at low levels of consumption consumers prefer to specialize, but that at high levels diversification becomes preferable. In other words, the indifference curves are concave at low levels of utility and convex at high levels of utility. Figure 3 depicts the indifference map. A consumer who devotes little time to art would patronize only one artist. A consumer who devotes more time, would patronize both artists. This result can be generalized: the more time one devotes to art, the larger would be his or her set of stars.

#### B. Pricing

So far I have assumed that the only cost to the consumer in the consumption of art is time. However, stardom produces savings in time costs and the star could absorb part of these savings.

Assume that the cost of production of art is zero. If there are pecuniary costs to the consumption of art, the budget constraint becomes

$$(6) \quad x(P_x + w + w/X) + y(P_y + w + w/Y) = Iw,$$

where  $P_x$  and  $P_y$  are the rental prices per unit of time of  $X$  and  $Y$ , respectively, and  $w$  is the wage rate. The consumer will choose  $X$  over  $Y$  as long as  $P_x - P_y < w/Y - w/X$ .<sup>4</sup>

#### C. Amateurs Who Excel

My model allows the star to remain a star even though an amateur could have a greater talent. To see this point, denote by  $U_1$  the utility from the consumption of  $X$ ,  $U_2$  the

<sup>3</sup>This paper has much in common with the literature on the bandwagon effect. It adds, however, to this literature by explaining the effect in one set of goods. Moreover, this set of goods is probably more prone to the bandwagon effect than other goods. One could imagine, for instance, a system without fashion in clothing. This would be the case where uniforms were required. Since, however, the source of the bandwagon effect in this paper is knowledge, the phenomenon of superstardom would be much more difficult to uproot in the goods discussed here.

<sup>4</sup>The  $X$  markup could be even larger, since the utility from an artist that consumers already know is greater. See the discussion of Figure 4 below.

## D. Different Tastes

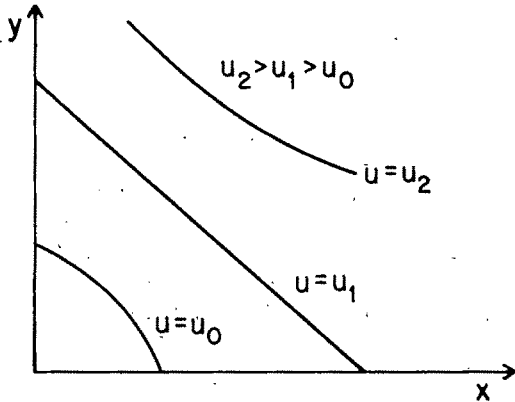


FIGURE 3

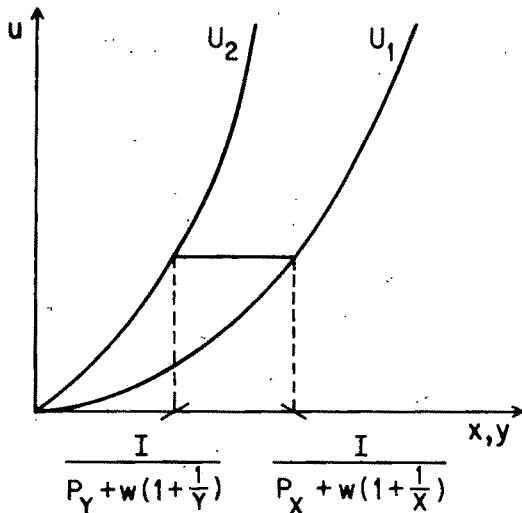


FIGURE 4

utility from the consumption of  $Y$ . If  $Y$  is more talented, then for  $x = y$ ,  $U_1(x) < U_2(y)$ . The quantities of  $x$  and  $y$  that the consumers could consume given his budget,  $I$ , are  $I/(P_X + w(1 + 1/X))$  and  $I/(P_Y + w(1 + 1/Y))$ , respectively. The consumer will continue to pick  $x$  as long as

$$(7) \quad U_1\left(I/P_X + w\left(1 + \frac{1}{X}\right)\right) > U_2\left(I/P_Y + w\left(1 + \frac{1}{Y}\right)\right).$$

Figure 4 depicts this condition.

Thus far I have assumed identical consumers and an identical unspecified artistic activity which they all consume. If consumers have different tastes, there would be different artistic activities (singing, painting, pottery) and within each activity different types of that activity (opera vs. rock music, abstract vs. realistic paintings). Consumers of each category have similar tastes, but this need not be the case across categories. Each category constitutes a market with its own stars. Of course, if there isn't any group of consumers with similar tastes there might not be any stars. But this is equally valid in the model developed here and in many other models, including Rosen's.

## E. Who Would Be the Star?

If everybody could be, who would be the star? My answer would be: luck would determine. (By luck, I mean factors other than talent.) But before I elaborate, a word on the relevance of the question to this paper.

According to this paper, stardom and money have similar characteristics. First, bills of all colors could serve as money and likewise all artists could be stars. Second, efficiency calls for only one money and likewise efficiency calls for very few artists with public recognition. Both characteristics exist in the case of money regardless of the process that determines which good would be the medium of exchange. I assert that the same independence exists here: the characteristics of stardom do not depend on the process by which a star evolves. Bearing this qualification in mind, the literature on the development of the medium of exchange, especially Robert Jones (1976), suggests a tentative outline of an answer, and is applied presently with some modifications.

Assume that at first consumers believe that all artists are equally likely to become stars, and that each consumer picks one artist at random. Assume that consumers live  $n$  periods and revise their prior distributions after each period. If there were a slight majority of consumers that picked  $X$  as their choice,  $X$  would snowball into the star because after each period this majority would increase. In

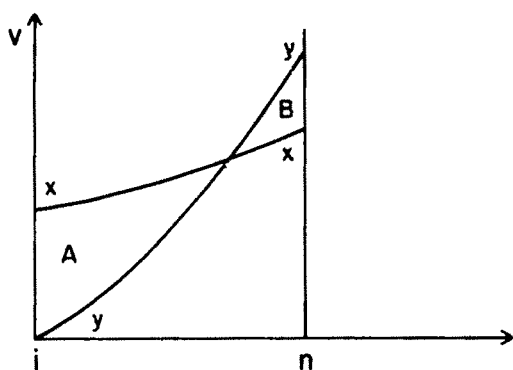


FIGURE 5

other words, if at any period of time an artist had a market share only slightly larger than everybody else, this share would increase steadily. Nonetheless, the lucky artist need not capture the whole market even if a time would come when everybody would recognize him or her as the star. A consumer who did not find whom the star was for a long time might be better off continuing with an erroneous choice than switching. He or she would be only beginning to learn about the star, whereas his or her knowledge about the other artist (who is equally talented) is already extensive. This advantage in knowledge about the nonstar has to be weighed against lower search costs for discussants if one were to switch to the star.

Figure 5 depicts a consumer who thought that  $X$  was the star, but in period  $i$  finds out that he was wrong.  $V$  is the "indirect" utility function, or the utility the consumer derived at period  $i$ ,  $i < n$  from consuming  $I/(1+1/X)$  units of  $X$  where  $I$  is the amount of time the consumer devotes at each period to both art and search cost. (As was shown, the consumer will choose either  $x$  or  $y$ , not both.) Since knowledge increases utility, previous consumption is an argument of

the utility function:  $V = V(I/(1+1/X), \bar{x})$ , where  $\bar{x}$  is the prior consumption of art. The terms  $\bar{x}$  and  $X$  could be replaced by  $\bar{y}$  and  $Y$ .

At period  $i$ , when the consumer discovers who the true star is, his indirect utility would be higher if he consumed the "false star" (the  $xx$  curve in Figure 5). However, since his search costs for discussants would be lower with the true star,  $V$  would grow faster if he switched (the  $yy$  curve). Only if area  $B$  is greater than  $A$  in Figure 5, will it pay to switch.

#### IV. Conclusion

This paper explains why a hierarchy in income could exist without a hierarchy in talent. In other words, it explains why there could be stars among individuals known to have equal talents. The main argument was that the phenomenon of stardom exists where consumption requires knowledge. The acquisition of knowledge by a consumer involves discussion with other consumers, and a discussion is easier if all participants share common prior knowledge. If there are stars, that is, artists that everybody is familiar with, a consumer would be better off patronizing these stars even if their art is not superior to that of others.

#### REFERENCES

- Jones, Robert A., "The Origin and Development of Media of Exchange," *Journal of Political Economy*, August 1976, 84, 757-75.
- Rosen, Sherwin, "The Economics of Superstars," *American Economic Review*, December 1981, 71, 845-58.
- Stigler, George and Becker, Gary, "De Gustibus Non Est Disputandum," *American Economic Review*, March 1977, 67, 76-90.



# Advertising and Economic Welfare

By LEN M. NICHOLS\*

Is advertising socially excessive? Economists have been reluctant to consider this question, for there appeared to be no rigorous method of analyzing cases in which advertising influences tastes. Some recent studies have made fundamental contributions to this debate (Avinash Dixit and Victor Norman, 1978; Yehuda Kotowitz and Frank Mathewson, 1979), but no consensus has emerged.

Dixit and Norman were the first to establish a rigorous methodology. They evaluated social welfare ( $SW$  = consumer's surplus + profits) under pre- and post-advertising market equilibria, using pre-advertising tastes for one ranking and post-advertising tastes for another ranking. When the  $SW$  rankings were identical, Dixit and Norman claimed that unambiguous inferences could be drawn. In general, they concluded that the profit-maximizing level of advertising will be excessive under a variety of product market structures.

Franklin Fisher and John McGowan (1979) criticized Dixit and Norman for including advertising in the utility function but ignoring the direct effect that advertising has on utility in their welfare analysis. Dixit and Norman (1979) countered that advertising merely shifts the preference ordering over goods, and that it is this preference yardstick that matters.

Welfare conclusions about advertising can be generated less controversially within the framework suggested by George Stigler and Gary Becker (1977). They treat utility as if it were generated by "commodities" which individuals produce with purchased

goods, time, human capital, and firms' advertising. Their commodities can be interpreted similarly to what Kevin Lancaster (1966) calls a "characteristic," the terminology used herein.<sup>1</sup>

As an example, consider tennis. We buy racquets and balls, not to own them per se, but rather to enjoy playing the game. Playing the game actually generates utility, and this characteristic is produced by an individual's racquet, balls, human capital (skill level), and time.

Thus preferences are ordered over characteristics, not over goods. Advertising can lower the shadow price of a characteristic by increasing the ability of the purchased good to produce the desired characteristic. This is accomplished by providing the consumer with different information or beliefs about the product than previously held.

Suppose a perennial tennis champion states a preference for a particular brand of racquet. This enhances the self-image of amateur owners of that brand and increases their enjoyment of the game without changing their expenditures. The per unit cost of the characteristic will decline, making current owners of the racquet likely repeat customers and current shoppers more likely to choose the star's preferred brand. Observe that advertising has increased the demand for a market good that produces the now cheaper char-

\*Assistant Professor of Economics, Wellesley College, Wellesley, MA 02181. I am indebted to Ralph Bradburd, David Lindauer, Ken Koford, Robert Feinberg, Ann Witte, and two anonymous referees for helpful comments and suggestions on earlier drafts. Any remaining errors and ambiguities are my sole responsibility.

<sup>1</sup>Lancaster developed an approach to consumer theory wherein market goods are desirable only because they provide utility-generating characteristics. In general, each market good provides many such characteristics, thus each input into the consumer's process of generating utility has many outputs. Stigler and Becker have in mind a world wherein each output, a single utility-generating commodity, is produced by multiple inputs; for example, purchased goods, human capital, time, and beliefs. A fully general model would obviously have multiple inputs and outputs. In that sense the approach employed herein is a special case, but one that is particularly well suited for the welfare analysis of advertising.

acteristic. The Stigler-Becker framework enables us to interpret an apparent *shift* in the product demand curve as a movement *along* a characteristic demand curve induced by increased advertising's reduction in the shadow price of the characteristic. Thus we can measure consumer's surplus as the area under the stable characteristic demand curve above the shadow price, and conduct welfare analysis of market equilibria under various levels of advertising without concern for which product demand curve is appropriate.

Kotowitz and Mathewson implicitly use the Stigler-Becker framework when they assume utility depends on two-dimensional quality. Advertising affects one of the dimensions, and the costlessly observable characteristics of the good affect the other. Thus utility is produced by the verifiable good and beliefs about that good, which are in turn affected by advertising. Their welfare conclusions about the social optimality of advertising are ambiguous, but they do isolate the factors that determine this result.

This paper generates welfare conclusions about advertising within the Stigler-Becker framework. A major result is that the social optimality of the profit-maximizing level of advertising depends upon the market structure of the produced characteristic and not upon the market structure of the purchased good. In addition, the method of analysis casts light upon the important insight of Dixit and Norman. Finally, the analysis produces a testable hypothesis on advertising's optimality for any firm or industry.

### I. Assumptions and the Model

Since profits are assumed to be maximized by firms' advertising choices, the effect of advertising on social welfare may be inferred by differentiating consumer's surplus with respect to advertising and evaluating this derivative at the profit-maximizing level of advertising.<sup>2</sup> This procedure is equivalent to

<sup>2</sup> This may be done unambiguously only when advertising by one firm does not affect the profits of another firm, i.e., there is no public good nor rival message cancellation effect of any firms' advertising. Future work will extend the analysis to these cases.

asking the question, could consumers be made better off with a different level of advertising than profit-maximizing firms will choose to provide?

This can be answered directly with the following model. Utility of the typical consumer is derived from the produced characteristics  $Z$  and  $Y$ . Following Stigler and Becker's simplification,  $Z$  is the product of a function of advertising and the quantity of a marketed good  $X$ ;  $Z = g(A)X$ . The function  $g(A)$ , the marginal characteristic productivity of  $X$ ,  $\partial Z / \partial X$ , summarizes the consumer's attitudes about the effectiveness of advertising. Ultimately, the consumer decides if a good is search or experience, and if an advertisement is informative or persuasive.<sup>3</sup> The  $Y$  is the composite nonadvertised good. Thus  $U = (g(A)X, Y)$ .<sup>4</sup>

Utility maximization subject to the budget constraint yields the following necessary conditions ( $U_z = \partial U / \partial Z$ ,  $U_y = \partial U / \partial Y$ ):

- (1)  $U_z / \lambda = P_x / g(A)$
- (2)  $U_y / \lambda = P_y$
- (3)  $I = P_x X + P_y Y$ ,

where  $\lambda$  is the Lagrangian multiplier, the marginal utility of income. Expression (1) shows that the shadow price of the characteristic  $Z$  is  $P_z = P_x / g(A)$ . This seems similar to Philip Nelson's (1974) notion of price per unit of utility.

Denoting the solutions to (1), (2), and (3) as  $X^*$ ,  $Y^*$ , and  $\lambda^*$ , we know that utility is maximized for a given level of advertising

<sup>3</sup> Past advertising that is now perceived to have been misleading will lower the characteristic productivity of currently produced (or remaining stocks of)  $X$ . Profit-maximizing firms will consider this future credibility cost when formulating their advertising messages. Kotowitz and Mathewson show that misleading advertising may be profit maximizing for a while given certain quality production cost and initial consumer perception parameters. If this is the case, expected characteristic flow will adjust and purchase decisions will be changed accordingly. This adjustment process is implicitly assumed to be performed each period.

<sup>4</sup> For all of the results to hold, we must make the conventional assumption that  $U(Z, Y)$  is strictly quasi concave.

when

$$(4) \quad U = U(g(A)X^*, Y^*).$$

This implies that income must be allocated according to

$$(5) \quad I = P_x X^* + P_y Y^*.$$

Differentiating (5) with respect to advertising yields a result that will be useful later in the welfare analysis,

$$(6) \quad -X^*(\partial P_x / \partial A) = P_x(\partial X^* / \partial A) + P_y(\partial Y^* / \partial A).$$

Now, to consider the question of consumer welfare and advertising, we differentiate (4):

$$dU/dA = U_x[X^*g'(A) + g(A)(\partial X^* / \partial A)] + U_y(\partial Y^* / \partial A).$$

Substituting from the first-order conditions, we have

$$dU/dA = \lambda \left[ (P_x X^* g'(A) / g(A)) + P_x(\partial X^* / \partial A) + P_y(\partial Y^* / \partial A) \right],$$

and using (6),

$$(dU/dA)/\lambda = [P_x X^* / A] \left[ (g'(A)A) / g(A) - (\partial P_x / \partial A)(A/P_x) \right].$$

Since  $(g'(A)A)/g(A)$  is the advertising elasticity of the characteristic productivity of  $X$ ,  $(E_{g,A})$ , and  $(\partial P_x / \partial A)(A/P_x)$  is the advertising elasticity of the price of  $X$ ,  $(E_{P_x,A})$ , we interpret

$$(7) \quad (dU/dA)/\lambda = (P_x X^* / A) [E_{g,A} - E_{P_x,A}].$$

Since  $\lambda$  is the marginal utility of income, the left-hand side of (7) may be interpreted as the consumer's willingness to pay for more advertising at the margin, the shadow price of advertising. If this willingness to pay is

positive (negative, zero) at the profit-maximizing level of advertising ( $\partial \pi / \partial A = 0$ ,  $A = A^*$ ), it may be concluded that advertising is under- (over-, optimally) supplied by the free market. Advertising is undersupplied if, at the profit-maximizing margin, increased advertising raises the characteristic productivity of the purchased good  $X$  more (in percentage terms) than it increases the market price of  $X$ . Advertising is oversupplied when the market price of  $X$  is increased proportionately more than the characteristic productivity is enhanced. Finally, equation (7) suggests that advertising is optimally supplied when advertising raises the market price and the characteristic productivity in equal proportions.

To focus the welfare analysis on the unchanging characteristic demand curve in  $P_z, Z$  space, note that  $P_z = P_x / g(A)$ ,  $\Rightarrow E_{P_z,A} = (\partial P_z / \partial A)(A/P_z) = E_{P_x,A} - E_{g,A}$ . Thus equation (7) could be written as

$$(8) \quad (dU/dA)/\lambda = (P_x X^* / A) [-E_{P_z,A}].$$

We can then say that advertising is undersupplied at the margin when it decreases  $P_z$ —the last unit of advertising allows the consumer to slide down the characteristic demand curve so that  $Z$  consumption and utility rise. If advertising raises  $P_z$ ,  $Z$  consumption and utility fall with the last unit of advertising, and advertising has been oversupplied. If advertising does not affect  $P_z$ ,  $E_{P_z,A} = 0$  and advertising is optimally supplied. In this case, the consumer has obtained the lowest possible  $P_z$ . The inability of advertising on  $X$  to affect the price of  $Z$  is how Stigler and Becker define a perfectly competitive characteristic market. Thus perfect competition in  $Z$  is a sufficient condition for the social optimality of advertising on  $X$ . The following analysis makes clear that the market structure of the purchased good  $X$  does not affect the social optimality of advertising on  $X$ .

## II. The Roles of the Firm and each Market's Structure

As in Stigler and Becker's analysis, the  $X$  producing firm is assumed to maximize profits by choosing output and advertising. In

general, profit maximization requires the marginal revenue of both output and advertising to equal their respective marginal costs. Despite the presence of advertising, the traditional pricing rules continue to hold.<sup>5</sup> When the market for  $X$  is perfectly competitive,  $P_x = MC_x$ , and when  $X$  is imperfectly competitive,  $P_x = MC_x / (1 - (1/E_{x,P_x}))$ .

The market for  $Z$  is more complicated, because consumers must produce  $Z$  with purchased goods and their own attitudes. In general, the same quantity of  $Z$  can be produced by a variety of  $X_i$  and  $g(A_i)$  combinations, where  $i$  could range over competing brands as well as functionally different products. A perfectly competitive  $Z$  market has an infinite variety of  $X_i$  and  $g(A_i)$  combinations that can produce the same  $Z$ , thus the equilibrium price of  $Z$  is exogenous to the individual firm producing and advertising  $X_i$ . No individual producer of an  $X_i$  can affect this  $P_z$  by advertising more or less, or by changing its product price, for a change in the price of  $Z$  indirectly produced by one firm will cause all consumers to shift to the lowest-cost  $Z$  technology. An imperfectly competitive  $Z$  market is one where perfect substitutes do not or are not perceived to exist, and in this case the price of  $Z$  may be altered by the advertising or pricing policies of a given  $X_i$  firm.

Four market structure configurations are possible, all permutations of perfect and imperfect  $X$  and  $Z$ . These four cases can be reduced to two analytically distinct ones: (I) imperfect  $X$  and perfect  $Z$ , and (II) perfect  $X$  and imperfect  $Z$ .<sup>6</sup>

*Case I:* The market for  $Z$  is perfectly competitive, but the market for  $X$  is not.

We know two facts about equilibrium: (a)  $P_x = MC_x / (1 - (1/E_{x,P_x}))$ , and (b) one firm varying the level of advertising on  $X$  cannot affect the perfectly competitive characteristic price, that is,  $(\partial P_z / \partial A) = 0$ .<sup>7</sup> Fact b

guarantees that  $E_{P_z,A} = 0 \Leftrightarrow E_{g,A} - E_{P_x,A} = 0$ , and thus in equations (7) and (8),  $(dU/dA)/\lambda = 0$ . We now have our welfare result that the level of advertising chosen by a profit-maximizing imperfect competitor in  $X$  is socially optimal if that  $X$  is used by consumers to produce a perfectly competitive characteristic. To expand on the example used by Stigler and Becker, in the social prestige characteristic market, ostensibly oligopolistic jewelry designers and furriers who indirectly compete with each other and other indirect producers of social prestige are compelled to advertise the social welfare-maximizing amount if the prestige market is perfectly competitive. The reason this result obtains is the existence of perfect substitutes in use.

*Case II:*  $X$  is perfectly competitive, but  $Z$  is not.

Although it may seem odd at first glance, Stigler and Becker point out that firms producing in perfectly competitive  $X$  markets will find it profitable to advertise if there are alternative technologies for producing  $Z$ , or if there is imperfect information about those technologies.<sup>8</sup> When  $Z$  is imperfectly competitive, advertising for  $X$  can affect the price of  $Z$ . From the consumer's first-order conditions,

$$\begin{aligned} P_z &= P_x / g(A) \Rightarrow P_x = P_z g(A), \\ &\Rightarrow (\partial P_x / \partial A) \\ &= (\partial P_z / \partial A) g(A) + P_z g'(A) \\ &\Rightarrow P_z = (\partial P_x / \partial A) (1/g'(A)) \\ &\quad - (\partial P_z / \partial A) (g(A)/g'(A)) \\ &\Rightarrow P_x = (\partial P_x / \partial A) (g(A)/g'(A)) \\ &\quad - (\partial P_z / \partial A) (\{g(A)\}^2 / g'(A)) = MC_x. \end{aligned}$$

<sup>5</sup>See Stigler and Becker, p. 85.

<sup>6</sup>The results for perfect  $X$  and perfect  $Z$  follow *a fortiori* from Case I. Similarly, the results for imperfect  $X$  and imperfect  $Z$  follow from Case II. All derivations are available from the author on request.

<sup>7</sup>See Stigler and Becker, p. 85.

<sup>8</sup>The implicit assumption in Stigler-Becker and here is that expanding output alone will not affect price, while advertising will. The  $X$  demand curve facing the firm could be infinitely elastic even though advertising can shift it upward within the limits set by the  $g(A)$  function and the profit-maximizing requirement that  $MR_A = MC_A$ .

Using the facts that  $(g(A)/g'(A)) = A/E_{g,A}$  and  $g(A) = P_x/P_z$ , we have

$$(\partial P_x / \partial A)(A/E_{g,A}) - (\partial P_z / \partial A)(P_x g(A)/P_z g'(A)) = MC_x$$

which, when divided by  $P_x$  and rearranged, yields

$$(9) \quad E_{P_x,A}/E_{g,A} = 1 + E_{P_z,A} [g(A)/g'(A)A].$$

Thus, perfection in the  $X$  market does not guarantee socially optimal advertising. Indeed, the result states that the level of advertising that maximizes profits also maximizes social welfare if and only if  $P_z$ , the shadow price of the utility generating characteristic, remains unchanged.<sup>9</sup>

This interpretation suggests a testable proposition on the social optimality of advertising. Since the characteristic  $Z$  is the ultimate object of utility, the demand for  $X$  is a derived demand. When advertising increases the characteristic productivity of  $X$  more than the price of  $X$ , utility maximizers will consume more of  $X$ . Thus if we estimate a sales (physical quantity) equation like

$$(10) \quad Q_s = \alpha + \beta W + \gamma A + \epsilon,$$

where  $W$  is a vector of all other causal influences on sales, a statistically significant  $\hat{\gamma} > 0$  implies that  $E_{P_z,A} < 0$ , and we may infer that advertising is being undersupplied.

Certainly this is more likely when advertising enhances competition, increases the effective price elasticity and thus lowers the market price of a traded good ( $E_{P_x,A} < 0$ ), as in the case of eyeglasses (Lee Benham, 1972). But note that advertising could *increase* the price of the market good and still be undersupplied. This is not possible in the Dixit and Norman formulation, for they assume that preferences are defined over goods, not characteristics. Their assumption that advertising lowers demand elasticity and

raises the price of the argument in the utility function is equivalent to assuming  $\partial P_z / \partial A > 0$ , that is, Dixit and Norman implicitly assume that advertising by a monopolist or oligopolist seller of  $X$  raises the value (characteristic productivity) of  $X$  less than the price of  $X$ . No consumer would choose this and, therefore, advertising is in their view generally oversupplied.

My analysis has identified a special case in which advertising could be oversupplied, specifically when the price of  $X$  is increased proportionately more than the characteristic producing capacity. This could be identified empirically by  $\hat{\gamma} < 0$  in equation (10). Advertising at the profit-maximizing margin reduces total quantity sold.

To illustrate this special case, consider adolescent-sized designer jeans. Marginal advertising messages on these physically similar products raise price more than the jeans' status productivity, but the lack of substitute  $Z$  technologies allow the indirect producers of  $Z$  to extract consumer's surplus. Advertising shifts the product demand curve to the right and makes it less elastic, allowing each firm to raise its price until equilibrium market quantity falls. The decline in equilibrium quantity brought about by advertising here is consistent with Dixit and Norman's analysis in which an increase in output was a necessary condition for  $SW$  to be improved by advertising.

Interestingly, the model suggests that we may interpret  $\hat{\gamma} = 0$  as evidence of optimal advertising. When this is observed, we know that advertising has enhanced the characteristic productivity as much as the market price.

### III. Summary and Conclusions

This paper has used the conceptual insight of Stigler and Becker to develop a framework for determining the social optimality of advertising. Thus it extends recent work by Dixit-Norman and Kotowitz-Mathewson.

One conclusion is that the social welfare-maximizing amount of advertising does not depend upon the market structure of the advertised good. Instead, the market structure of the utility-generating characteristic is crucial. The profit-maximizing level of advertising maximizes social welfare if the char-

<sup>9</sup>Unchanged  $P_z \Rightarrow E_{P_z,A} = 0 \Rightarrow E_{P_x,A} = E_{g,A} \Rightarrow (dU/dA)/\lambda = 0$ .

acteristic productivity of the advertised good is raised as much (proportionately) by advertising at the margin as is the price of the advertised good. This could be inferred when the observed marginal impact of advertising on the quantity of sales is zero. A positive impact of advertising on sales can be taken as evidence of too little advertising—consumers are willing to pay for more. An observed negative impact of advertising on sales suggests that advertising is being oversupplied.

More complicated models should be developed for many remaining issues. I abstracted away from multiproduct firms, where there may be advertising spillover effects across products of a well-known firm. Similarly, the externalities of brand-specific advertising on the characteristic productivities of generic products were ignored. Another interesting complication is that of a single market good being capable of producing multiple characteristics. Finally, it seems that the analysis of advertising in this paper would apply in general to most other forms of product differentiation.<sup>10</sup>

Thus, the model should be interpreted as a polar extreme, not yet ready for definitive policy recommendations. Still, the promise of this characteristic method of analyzing advertising and economic welfare is clear. It

can provide straightforward conclusions in cases where advertising changes preferences over marketed goods, and it can generate clearly testable hypotheses.

## REFERENCES

- Benham, Lee, "The Effect of Advertising on the Price of Eyeglasses," *Journal of Law & Economics*, October 1972, 15, 337-52.
- Dixit, Avinash K. and Norman, Victor D., "Advertising and Welfare," *Bell Journal of Economics*, Spring 1978, 9, 1-17.
- and ———, "Advertising and Welfare: Reply," *Bell Journal of Economics*, Autumn 1979, 10, 728-29.
- Fisher, Franklin M. and McGowan, John J., "Advertising and Welfare: Comment," *Bell Journal of Economics*, Autumn 1979, 10, 726-27.
- Kotowitz, Yehuda and Mathewson, Frank, "Advertising, Consumer Information, and Product Quality," *Bell Journal of Economics*, Autumn 1979, 10, 566-88.
- Lancaster, Kevin J., "A New Approach to Consumer Theory," *Journal of Political Economy*, January/February 1966, 74, 132-57.
- Nelson, Philip, "Advertising as Information," *Journal of Political Economy*, July/August 1974, 82, 729-54.
- Stigler, George J. and Becker, Gary S., "De Gustibus Non Est Disputandum," *American Economic Review*, March 1977, 67, 76-90.

<sup>10</sup>I am indebted to an anonymous referee for this observation.

# Oligopoly and the Incentive for Horizontal Merger

By MARTIN K. PERRY AND ROBERT H. PORTER\*

Economists have articulated the welfare implications of horizontal mergers.<sup>1</sup> Mergers can reduce industry competition and so result in higher prices for consumers. But, on the other hand, mergers may give rise to efficiency gains (for example, scale economies) that reduce the cost of production or distribution. These welfare discussions of merger often implicitly assume that firms always have an incentive to merge, that is, they can increase profits by raising price and/or reducing cost. This assumption derives from at least two sources. First, at an empirical level, the wave of merger activity at the turn of the century is thought to have been undermined by the Sherman and Clayton Acts, rather than a lack of profitable opportunities. Similarly, the periodic merger waves since then have often been attributed to reduced antitrust enforcement activities. Second, at a theoretical level, industry models such as the symmetric Cournot model exhibit higher prices and profit per firm as the number of firms in the industry is reduced.<sup>2</sup> Thus, there would seem to be profit gains from merger which reduces the number of firms.

However, George Stigler (1950) and others have argued that firms which do not participate in a merger may benefit more than the participants. When a merger occurs, the new firm will typically reduce its production below the combined output of its constituent

firms. As a result, industry price will increase. (This assumes that the cost reductions associated with merger are not too large.) Nonparticipants will then expand output and profit from the higher industry price. Thus, merger participants do not capture all the profits that result from their merger. Because of this externality, mergers which would increase total industry profits need not be privately profitable.

In their 1983 paper, Stephen Salant, Sheldon Switzer, and Robert Reynolds (S-S-R) reexamined the incentive to merge in a static framework. They employed a symmetric Cournot model with linear final demand, and identical constant average costs for each of  $n$  firms. Even though profits per firm are higher with fewer firms, S-S-R demonstrated that the profits of one firm in an  $n$ -firm oligopoly are lower than the profits of two firms in an  $(n+1)$ -firm oligopoly. (This is also true if there are fixed costs that are not large.) From this, they concluded that merger is generally unprofitable in a Cournot oligopoly. The exception to this finding occurs when duopolists merge into monopoly.

However, the S-S-R model severely understates the incentive to merge. The problem is that mergers are not well-defined conceptually. Merger of two firms from a symmetric equilibrium of  $(n+1)$  firms should result in an equilibrium with  $(n-1)$  old firms and one new firm that is "larger" than the others. In particular, the new firm should have access to the combined productive capacity of both merger partners. In the S-S-R model, the merged firm does not differ from the others; it continues to have access to the same technology. Thus, rather than finding that merger is unprofitable, S-S-R find that "lock-up" is unprofitable. The so-called merger increases price to the external benefit of all firms, but the new firm foregoes the production and profits of one of the two original firms. This could be profitable only when an industry monopoly is created from a duopoly. Models

\*Bell Communications Research, 600 Mountain Avenue, Murray Hill, NJ 07974, and Department of Economics, State University of New York, Stony Brook, NY 11794, respectively. We have benefited from the comments of John Panzar. This work was initiated during 1983 while Porter was a postdoctoral fellow at Bell Laboratories on leave from the University of Minnesota. This article represents our views and assumptions and not necessarily those of Bell Communications Research or the Bell Operating Companies.

<sup>1</sup>Oliver Williamson (1968) documents these effects.

<sup>2</sup>For example, see Perry (1984) and Jesus Seade (1980).

with constant average cost invite this conceptual fallacy by obstructing any notion of assets or firm size.

Raymond Deneckere and Carl Davidson (1983) have shown that the existence of product differentiation can reverse the S-S-R results where the merged firm continues to produce all the products of its constituent firms. We instead propose to deal with the notion of merger directly from the cost side. Unlike S-S-R we can find many circumstances in which an incentive to merge exists, even though the product is homogeneous. Moreover, we can identify the behavioral and structural characteristics that will give rise to such an incentive.

Our formulation of mergers has the following advantages. We explicitly specify a tangible asset that the merged firm acquires from its two partners, which increases the output it can produce at a given average cost. Assets are captured by assuming a capital factor that is in fixed supply for the industry. This structure allows us to address the industry asymmetries caused by the merger of subsets of firms. A merged firm faces a different maximization problem because of its altered cost function and new strategic considerations. Finally, this notion of assets is uniform and fungible enough to discuss any combination of firms in a meaningful and convenient way.

In Section I, we define the cost and demand structure and in Section II, examine a dominant oligopoly model. There is a competitive fringe from which mergers occur to create oligopolists of a particular size. In Section III, we eliminate the fringe and examine a model in which there are large and small oligopolists. Large oligopolists are formed by the merger of two small oligopolists. For both models we characterize the circumstances under which there exists an incentive to merge.

### I. Cost and Demand Structure

The cost structure is the key feature of the model. We assume that there is a factor whose total supply is fixed to the industry, say capital. It is a necessary input in the production process. What will distinguish

firms is the amount of capital which they own. For convenience, we normalize the total quantity of capital to be unity.<sup>3</sup>

The cost function of a firm that owns a fraction  $s$  of the capital stock and produces output  $x$  is denoted  $C(x, s)$ . The output  $x$  is produced with a combination of the fixed factor and a vector of variable inputs  $z$ , according to a smooth concave production function,  $x = F(z, s)$ . Then  $C(x, s)$  is dual to  $F(z, s)$ , where the cost function implicitly subsumes the factor prices corresponding to the input vector  $z$ . We further assume that  $F$  is linearly homogeneous in  $z$  and  $s$ , so that there are constant returns to scale. This implies that the cost function will be linearly homogeneous in  $x$  and  $s$ . Thus, any proportionate increase in output and capital will result in an equiproportionate increase in production costs. Because of the presence of a fixed factor of production, the marginal cost function of any single firm is increasing.<sup>4</sup> Furthermore, since  $C_1(x, s)$  is homogeneous of degree zero, the competitive industry supply curve is fixed at  $C_1(x, 1)$ , irrespective of the distribution of capital among firms. Thus, the competitive industry supply curve is rising, and the marginal cost function of each firm mirrors this industry marginal cost function. In particular, a firm's marginal cost curve is obtained by horizontally shifting the industry marginal cost curve inward, in pro-

<sup>3</sup> Obviously, this suppresses de novo entry into the industry. With entry, the gains to merger diminish in that the magnitude of the price increases would be limited by the supply of new competitors. To capture this effect, it might be preferable to specify a rising supply curve for the fixed factor. We could then examine the impact of entry, which becomes increasingly costly, on the ability of existing firms to raise price through merger and therefore on their reduced incentive to merge. However, this would substantially complicate the model with little gain in insight. The fixed-factor assumption is a limiting case of upward-sloping factor supply. Its convenience derives from fixing the size of the industry so that we can readily examine changes in the number of firms and their sizes.

<sup>4</sup> If  $C(x, s)$  is linearly homogeneous,  $C_i(x, s)$  will be homogeneous of degree zero, where  $C_i(\cdot)$  is the derivative of  $C(\cdot)$  with respect to its  $i$ th argument. Then  $C_{12}(x, s) < 0 < C_{11}(x, s)$ , i.e., marginal costs decrease as capital increases, and so marginal costs must increase with output levels. (Recall, from Euler's Theorem, that  $xC_{11}(x, s) + sC_{12}(x, s) = 0$ .)



portion to the firm's share  $s$  of the industry capital stock. Thus, we can reconstitute the industry structure in any desired pattern without altering the overall size of the industry.<sup>5</sup>

Since the cost function is linearly homogeneous in  $x$  and  $s$ , we specifically rule out scale economies as a motive for merger. Rather, we wish to focus on the incentives to merge that arise solely from firm size and behavior in an imperfectly competitive environment.

In order to illustrate the gains and losses associated with merger, we specify particular cost and demand functions. Both demand and marginal cost are assumed to be linear functions of output. The cost function for a firm with capital stock  $s$  is

$$(1) \quad C(x, s) = s \cdot g + d \cdot x + (e/(2 \cdot s)) \cdot x^2.$$

Clearly,  $C(x, s)$  is linearly homogeneous. Industry fixed costs,  $g$ , are distributed in proportion to holdings of capital. The resulting marginal cost function is linearly increasing with an intercept of  $d$ :

$$(2) \quad C_1(x, s) = d + (e/s) \cdot x.$$

The marginal cost function rotates about the intercept as the amount of capital,  $s$ , increases or decreases.

Similarly, we assume that the industry inverse demand function is linear:

$$(3) \quad P(Z) = a - b \cdot Z,$$

where  $Z$  is industry output. In the next section, we discuss mergers in an oligopoly model with a competitive fringe.

## II. Oligopoly with a Competitive Fringe

In this section, we consider an industry with  $n$  oligopolists, each of which own a fraction  $s$  of the industry capital, and a

competitive fringe which owns the remaining capital,  $1 - ns$ .<sup>6</sup> We then examine the incentives of some of the fringe firms to coalesce and form a new oligopolist with capital share  $s$ . Enough capital must be held by the fringe for this to be feasible, so that  $s \leq 1 - sn$ . If this inequality is strict, then the fringe will contract but not vanish when the merger occurs.

Let  $V$  be the quantity supplied by the fringe firms, and  $X$  the total quantity supplied by the oligopolists. Fringe firms produce at output levels which equate price and marginal costs. With the horizontal summation property of the marginal cost function, this fringe equilibrium condition can be expressed simply as

$$(4) \quad P(X + V) = C_1(V, 1 - sn).$$

The quantity supplied by the fringe is implicitly defined by (4) as a function of the output of the oligopolist,  $V(X)$ .

The inverse residual demand function facing the oligopolists is then  $P(X + V(X))$ . They behave as a Stackelberg group with respect to the competitive fringe. This is a simple generalization of the dominant firm model to dominant oligopoly. The equilibrium condition for the oligopolists equates perceived marginal revenue to marginal cost, given a capital stock  $s$  for each.

The imperfectly competitive interaction of the oligopolists will be summarized by a conjectural variation. Let  $\delta$  be the conjectured output response of the other oligopolists to a unit change in own output. If  $\delta = -1$ , then the oligopolists behave competitively. Therefore, we assume  $\delta > -1$ . If  $\delta = 0$ , we have the Cournot model in which each oligopolist assumes that the others will not respond to output changes. Finally, if

<sup>5</sup>See Perry (1978) for a more complete derivation of this construction and a use of it to examine backward integration by a monopolist into a competitive upstream industry.

<sup>6</sup>The maximum size of an oligopolist could be thought of as arising from the Justice Department antitrust guidelines. (See *Merger Guidelines*, 1982.) For example, a guideline that triggered a challenge for mergers that put the industry over a 60 percent four-firm concentration ratio would imply  $s = .15$  under the assumption of symmetric oligopolists. Similarly, a Herfindahl index limit of 1600 would imply  $s = .20$  if there were four oligopolists, as  $4 \cdot (100 \cdot s)^2 = 1600$ .

$\delta = (n-1)$ , the  $n$  oligopolists act collusively to maximize joint profits. Since the fringe supply is constrained by its fixed capital stock, profitable opportunities remain for the oligopolists.

The symmetric oligopoly equilibrium output level is then given by

$$(5) \quad P + (1 + \delta) \cdot (1 + V') \cdot (X/n) \cdot P' \\ = C_1(X/n, s),$$

where  $P = P(X + V(X))$  and primes denote first derivatives. Given the cost and demand specifications of Section I, we can solve (4) and (5) to obtain the equilibrium output of the oligopolists and the fringe,

$$(6) \quad X(n) = \frac{(a-d)sn}{e + b[1 + s(1 + \delta)]},$$

$$(7) \quad V(n) =$$

$$\frac{(a-d)(1-sn)\{e + b[1 - sn + s(1 + \delta)]\}}{\{e + b[1 + s(1 + \delta)]\} \cdot [e + b(1 - sn)]}.$$

As a consequence of the linearity, the output produced by any single oligopolist,  $X(n)/n$ , is independent of  $n$ . Industry output  $Z(n)$  is the sum of  $X(n)$  and  $V(n)$ , so that

$$(8) \quad Z(n) = \frac{(a-d) \cdot K(n)}{e + b \cdot K(n)},$$

$$\text{where } K(n) = \frac{e + b(1 - sn)[1 + s(1 + \delta)]}{e + b[1 - sn + s(1 + \delta)]}.$$

Total output  $Z(n)$  is an increasing function  $K(n)$ , which is decreasing in  $n$ . Therefore, price rises as the number of oligopolists increases. This occurs because the competitive fringe becomes a smaller fraction of the industry, so that on balance the industry behaves less competitively. Without this price effect, there would be no incentive for fringe firms to merge. This occurs because a new oligopolist supplies less than did the compo-

nent share  $s$  of the fringe prior to the merger:

$$(9) \quad \frac{sV(n)}{1 - sn} - \frac{X(n+1)}{n+1} = \\ \frac{(a-d)s}{\{e + b[1 + s(1 + \delta)]\}} \cdot \frac{sb(1 + \delta)}{e + b(1 - sn)} > 0.$$

Therefore, the profits of the merged firm can exceed those of its constituent firms only if the merger results in a price rise sufficient to offset the lower output level.

In order to evaluate this incentive to merge, we compare the profits of fringe firms owning a fraction  $s$  of the capital stock,  $\pi_f(n)$ , when there are  $n$  oligopolists, with the profits of one of  $n+1$  oligopolists,  $\pi_0(n+1)$ . A fraction  $s/(1 - sn)$  of the fringe earns

$$(10) \quad \pi_f(n) = s/(1 - sn) \\ \cdot [P(Z(n)) \cdot V(n) - C(V(n), 1 - sn)],$$

while the merged firm would earn

$$(11) \quad \pi_0(n+1) = P(Z(n+1)) \\ \cdot \frac{X(n+1)}{n+1} - C\left(\frac{X(n+1)}{n+1}, s\right).$$

Since both entities incur fixed costs of  $s \cdot g$ ,  $g$  clearly has no impact on the incentive to merge. Moreover, we also discover that the comparison of profits is independent of the scaling parameters  $a$  and  $d$ . But otherwise, the incentive to merge depends upon the slope parameters  $b$  and  $e$ , and more importantly, the conjectural variation  $\delta$  and the number of firms  $n$ .

Consider the first merger. When a single oligopolist forms, we restrict  $\delta = 0$  since the oligopolist has no direct competitors. After substituting the appropriate expressions into (10) and (11), we find that  $\pi_0(1) > \pi_f(0)$  for all parameter values. Thus, there is always an incentive for a dominant firm to form from the fringe.

When  $n=1$ , a second merger could involve a movement in  $\delta$  from zero to a value in the interval  $(-1, 1)$ , as unity is the collu-

sive conjecture. We find that  $\pi_0(2) > \pi_f(1)$  if

$$(1-\delta)^2(e+b)^3 + 2(1+\delta^2)(e+b)^2bs \\ + (e+b)b^2s^2 + 2\delta b^3s^3 > 0.$$

This expression is positive for all nonnegative values of  $\delta$ . However, if  $\delta$  is sufficiently close to  $-1$ , this expression will be negative. Thus, a second firm will form from the fringe unless the resultant equilibrium is too competitive. In general, if oligopoly behavior becomes more competitive as a result of the merger, merger will be less profitable.

For  $n \geq 2$ , a comparison of equations (10) and (11) is computationally cumbersome if  $\delta$  changes with  $n$ . Thus, we hereafter assume that  $\delta$  remains unchanged after a merger occurs.

Consider now the third and subsequent mergers (obviously, we require  $s < 1/3$ ). Substituting from the functional specifications of Section I and the equilibrium quantities (6)–(8), we find that the incentive to merge can be simplified to

$$(12) \quad \pi_0(n+1) \gtrless \pi_f(n)$$

$$\text{as} \quad I(\delta, n) = -sb(1-\delta)n$$

$$+ [(1-\delta)(e+b) + sb(1+\delta)] \gtrless 0.$$

First, consider cases where  $\delta < 1$  (which includes the Cournot case where  $\delta = 0$ ). For these cases, the term  $I(\delta, n)$  is decreasing in  $n$ . Since  $I(\delta, n)$  is positive at the largest possible value of  $n$ ,  $(1/s)-1$  (where the next merger eliminates the fringe),  $I(\delta, n)$  must be positive for all relevant  $n$ . Also,  $I(1, n)$  is clearly positive. Therefore, for  $\delta \leq 1$ , there is always an incentive for an additional merger from the fringe.

For  $\delta > 1$ ,  $I(\delta, n)$  is increasing in  $n$ . Evaluating  $I(\delta, n)$  at  $n = 2$ , the smallest possible value of  $n$ , we find that it is positive for  $\delta < (e+b-sb)/(e+b-3sb)$ . For these  $\delta$ ,  $I(\delta, n)$  is positive, so there is an incentive to merge for all feasible values of  $n$  and  $\delta$  (recall that  $\delta \leq n-1$ ). Since  $(e+b-sb)/(e+b-2sb) > 1$ , this case subsumes the previous case of  $\delta \leq 1$ .

Relatively competitive behavior gives rise to an incentive to merge from the fringe. However, when  $\delta$  is less competitive, we obtain cases in which no such incentive exists. In particular, when  $I(\delta, n)$  is evaluated at the maximal value of  $n$ ,  $(1/s)-1$ , we see that there can never be an incentive to merge at any  $\delta > (e+2sb)/e$ , which also satisfies  $\delta \leq n-1$ . This leaves a range of intermediate values of  $\delta$  for which there is an incentive to merge when  $n$  is large but not when  $n$  is small. This arises because  $I(\delta, n)$  is increasing in  $n$ . We can summarize the results for  $n \geq 2$  as follows:

$$(13a) \quad \pi_0(n+1) > \pi_f(n)$$

$$\text{for} \quad \delta < (e+b-sb)/(e+b-3sb),$$

$$(13b) \quad \pi_0(n+1) \gtrless \pi_f(n)$$

$$\text{as} \quad n \gtrless \frac{(\delta-1)(e+b)-sb(1+\delta)}{sb(\delta-1)}$$

$$\text{for} \quad \frac{e+b-sb}{e+b-3sb} < \delta < \frac{e+2sb}{e} \text{ and } n \geq 1+\delta$$

$$(13c) \quad \pi_0(n+1) < \pi_f(n)$$

$$\text{for} \quad (e+2sb)/e < \delta \leq n-1.$$

The intuition for these results can best be understood by reference to Figure 1. Assume that  $b=1$ ,  $d=0$ , and  $e=s$ . Consider the incentive for the next to last merger from the fringe. The output  $OA$  is the equilibrium output of the  $n=(1/s)-2$  firms in the dominant oligopoly. Since output per firm is independent of  $n$  in this linear case,  $OA$  is also independent of the merger choice of the  $(n+1)st$  firm examined here. This leaves a residual demand curve of  $QT$  facing the fringe. The marginal cost curve for the fringe firms considering merger is  $AR$  ( $C_1(x, s) = ex/s = x$ ). This is also the supply curve of the fringe that would remain after such a merger. Thus, the residual demand facing the firms in the fringe considering merger is  $ET$ . Without merger, these firms would continue to behave competitively, produce  $BH$  at a



### III. Oligopoly of Small and Large Firms

Here we briefly consider an industry with  $n$  "large" oligopolists and  $m$  "small" oligopolists. The large oligopolists own  $s$  of the capital stock as in Section II. But instead of a competitive fringe, we assume that the remainder of the industry is composed of small oligopolists that own only  $s/2$  of the capital stock. The capital constraint  $s \cdot n + (s/2) \cdot m = 1$  must now hold. This structure provides a closer analogy to the S-S-R model in that we can examine the incentive for two small firms to merge into one large firm.

Let  $V$  be the quantity supplied by the small firms, and  $X$  the quantity supplied by the large firms. Despite their differing sizes, all firms now have the same conjectural variation  $\delta$ . Thus, the industry equilibrium can be defined by the symmetric first-order conditions for each firm size:

$$(15a) \quad P(X+V) + (1+\delta) \cdot V/m$$

$$\cdot P'(X+V) = C_1(V/m, s/2)$$

$$(15b) \quad P(X+V) + (1+\delta) \cdot X/n$$

$$\cdot P'(X+V) = C_1(X/n, s).$$

Given the cost and demand specifications of Section I, we can solve this system for the output of the two groups of firms as a function of the number of large firms  $n$ :

$$(16a) \quad X(n)$$

$$= \frac{(a-d) \cdot [b(1+\delta) + 2e/s] \cdot n}{\Delta(n)},$$

$$(16b) \quad V(n)$$

$$= \frac{2(a-d) \cdot [b(1+\delta) + e/s] \cdot [1/s - n]}{\Delta(n)},$$

where  $\Delta(n) = [b(1+\delta) + e/s] \cdot [b(1+\delta) + 2 \cdot (e+b)/s] - b^2(1+\delta) \cdot n$ . Industry output  $Z(n)$  is the sum of  $X(n)$  and  $V(n)$ , so that

$$(17) \quad Z(n)$$

$$= \frac{(a-d) \cdot [b(1+\delta)(2/s - n) + 2e/s^2]}{\Delta(n)}.$$

It is a simple matter to show that total output is a decreasing function of the number of large firms  $n$ . Thus, mergers again result in an increase in price to consumers. But unlike the previous model, firm behavior remains the same and price increases because there are now fewer firms in the industry.

The output of two small firms prior to merger is greater than output of the one postmerger firm. Thus, an incentive to merge requires that the increase in industry price be sufficient to offset the reduction in output of the merged firm. To evaluate this incentive, we simply compare the profits of one postmerger large firm,  $\pi_l(n+1)$ , with the profits of two premerger small firms,  $2 \cdot \pi_s(n)$ . The profits of each small firm before merger are

$$(18) \quad \pi_s(n) = P(Z(n))$$

$$\cdot \left[ \frac{s \cdot V(n)}{2 \cdot (1 - sn)} \right] - C \left( \frac{s \cdot V(n)}{2 \cdot (1 - sn)}, \frac{s}{2} \right),$$

while the profits of each large firm after merger are

$$(19) \quad \pi_l(n+1) = P(Z(n+1))$$

$$\cdot \frac{X(n+1)}{n+1} - C \left( \frac{X(n+1)}{n+1}, s \right).$$

Using the demand and cost assumptions of Section I and substituting from (15) and (16), we find that

$$(20) \quad \pi_l(n+1) \gtrless 2 \cdot \pi_s(n) \quad \text{as} \quad \Delta(n) \gtrless \bar{\Delta},$$

where  $\bar{\Delta} = (2b^2 \cdot q) / [2q - (q + e/s) \cdot (b(1+\delta) + q)^{1/2} \cdot q^{-1/2}]$  with  $q = b(1+\delta) + e/s$ . Since  $\Delta(n)$  is a decreasing function of  $n$ , three cases arise. First, if  $\max_n \Delta(n) = \Delta(0) < \bar{\Delta}$ , then there would always be an incentive to merge. Second, if  $\min_n \Delta(n) = \Delta((1/s) - 1) > \bar{\Delta}$  then there would never be an incentive to merge. And third, mixed cases can arise in which there is no incentive for the first mergers, but that if enough large firms already exist, there would be an incentive for the remaining small firms to merge.<sup>8</sup>

<sup>8</sup>Note that the three cases in this second model qualitatively resemble the three cases (13a)–(13c) in the

We can now illustrate that the S-S-R results generally fail to hold in our model with assets and two firm sizes. Let the conjectural variation be Cournot ( $\delta = 0$ ) as in the S-S-R model. If we then made the further S-S-R assumption that marginal costs were constant ( $e = 0$ ), conditions (15a) and (15b) would define an equilibrium with all firms being the same size. Thus, rising marginal costs ( $e > 0$ ) are necessary to define an equilibrium with differing firm sizes. Now recall that the S-S-R model yields no incentive for merger except when duopolists merge to form a monopolist. In our model when  $s = 1/2$  (small-firm duopoly), we do find that there is always an incentive to merge into one large firm.<sup>9</sup> However, our results depart considerably from those of S-S-R when  $s \leq 1/3$ , in particular, the incentive to merge often remains. When  $e > 3b$ ,  $\Delta(0) < \bar{\Delta}$  for all  $s \leq 1/3$  so that there is always an incentive to merge. Even when  $e < 3b$ , the incentive for the first merger (and thus subsequent mergers) still exists if there are few enough firms in the small oligopoly, that is, for  $s$  large enough. For  $e = 2b$ , there is an incentive for the first merger when  $s \geq 1/7$  (7 or less small firms), and no such incentive when  $s \leq 1/8$ . Similarly for  $e = b$ , there is an incentive when  $s = 1/2$  or  $1/3$ , but no incentive when  $s \leq 1/4$ . The exact S-S-R result only holds when  $e \leq b/2$ .<sup>10</sup> Note that  $e \leq b/2$  would encompass examples with  $e$  close to zero. Thus, even though the S-S-R model with  $e = 0$  has little theoretical appeal, the S-S-R results are not at variance with the limiting case of our model. However, we can generally conclude that there is a broad range of the parameters

for which there is an incentive for the first as well as all subsequent mergers.

#### IV. Conclusion

The incentive to merge depends upon a complex resolution of two forces. First, a merger results in a price increase. But second, the output of the merged firm declines relative to that of its partners prior to the merger. The price increase benefits all firms, but contrary to the results of S-S-R, it can often be sufficient to compensate for the output reduction of the merged firm and increase profits. Once one allows a merged firm to be twice as "large" as each partner, the output reduction is not nearly as severe as in the S-S-R model. However, we also find that there need not always be an incentive to merge. For the dominant oligopoly model, the conjectural variation alters this tradeoff. In particular, competitive conjectures are more conducive to merger. For the large-small oligopoly model, the incentive to merge also depends in a complex way upon demand and cost parameters.

Both models also provide a number of industry scenerios. In the dominant oligopoly model, we could find merger by the entire industry or by only the first or second firms. In the large-small oligopoly model, we could find no mergers at all. And in either model, we could find the unstable situation in which mergers would not occur unless the industry was sufficiently concentrated to begin with. Thus, even our simple models are very rich in their implications on the incentive for horizontal merger.

#### REFERENCES

- d'Aspremont et al., Claude, "On the Stability of Collusive Price Leadership," *Canadian Journal of Economics*, February 1983, 16, 17-25.
- Deneckere, Raymond and Davidson, Carl, "Coalition Formation in Noncooperative Oligopoly Models," unpublished, Michigan State University, 1983.
- Donsimoni, Marie-Paule, "Stable Heterogeneous Cartels," unpublished, Université Catholique de Louvain, 1983.

dominant oligopoly model. However, there is a quantitative difference in that the incentive to merge will generally be greater in the dominant oligopoly model. In that model, the  $m$  oligopolists do not expand output with a new merger; whereas the  $m$  large oligopolists in the large-small model do expand with another merger. Thus, the price increase is dampened in this second model.

<sup>9</sup>Specifically, we can easily show that  $\Delta(0) < \bar{\Delta}$  when  $\delta = 0$  and  $s = 1/2$ .

<sup>10</sup>These examples follow quickly once the relevant condition  $\Delta(0) < \bar{\Delta}$  has been simplified to  $-2b^3 - 8b^2/s - 17b^2e/s - 4b[5e^2 + 4e - 2b^2]/s^2 - 4e(e+b)(e-3b)/s^3 < 0$ . For further detail, see Section III of our 1983 working paper.

- \_\_\_\_\_, Economides, N. S., and Polemarchakis, H. M., "Stable Cartels," unpublished, Université Catholique de Louvain, 1983.
- Perry, Martin K., "Vertical Integration: The Monopsony Case," *American Economic Review*, September 1978, 68, 561-70.
- \_\_\_\_\_, "Scale Economies, Imperfect Competition, and Public Policy," *Journal of Industrial Economics*, March 1984, 32, 313-33.
- \_\_\_\_\_, and Porter, Robert H., "Oligopoly and the Incentive for Horizontal Merger," working paper, University of Minnesota, 1983.
- Salant, Stephen W., Switzer, Sheldon and Reynolds, Robert J., "Losses from Horizontal Merger: The Effects of an Exogenous Change in Industry Structure on Cournot-Nash Equilibrium," *Quarterly Journal of Economics*, May 1983, 98, 185-99.
- Seade, Jesus, "On the Effects of Entry," *Econometrica*, March 1980, 48, 479-89.
- Stigler, George J., "Monopoly and Oligopoly by Merger," *American Economic Review Proceedings*, May 1950, 40, 23-34.
- Williamson, Oliver, E., "Economies as an Antitrust Defense: Welfare Tradeoffs," *American Economic Review*, March 1968, 58, 18-36.
- U.S. Department of Justice, *Merger Guidelines*, 47 Fed. Reg. 28, 493, Washington: USGPO, 1982.



# Military Enlistments: What Can We Learn from Geographic Variation?

By CHARLES BROWN\*

Since the decision in 1973 to end the draft and return to reliance on voluntary enlistments, the number and quality of volunteers who could be attracted at politically realistic wages has been a key question. Because the armed forces adjust required qualifications when the supply of enlistees exceeds or falls short of the desired number of new recruits, the real policy issue is the availability of enlistees with desired characteristics such as a high school degree and high mental-test scores. The number of such high-quality enlistees is supply determined, but the total number of enlistees will be dominated (and perhaps determined) by recruiting targets.

It is difficult to summarize the results of previous econometric work on the supply of enlistees because the supply measure has been defined in so many ways. Some studies analyze total Defense Department enlistments, while others look at enlistments in particular services. Most focus on high-quality recruits, though "high-quality" is defined in different ways. Estimates of the elasticity of enlistments with respect to military compensation center a bit below 1.0 (Charles Dale and Curtis Gilroy, 1983; Richard Morey and John McCann, 1983). Elasticities with

respect to unemployment are much smaller; Morey and McCann's survey's range of 0.2 to 0.5 is representative. The strongest statement of the unimportance of unemployment is Colin Ash, Bernard Udis, and Robert McNown's conclusion that "The evidence on the lack of an unemployment effect on accessions is overwhelming" (1983, p. 147).

Most previous studies have used time-series data, and most of these pool pre- and post-1973 data in order to achieve an adequate sample size.<sup>1</sup> The impact of both the draft and the Vietnam war would be expected to influence supply, yet they are difficult to hold constant statistically. Because the Vietnam war years were a time of low military (relative to civilian) pay and low unemployment rates, inadequate control for the effect of the draft and the war are likely to bias estimates of the effects of military pay and civilian unemployment.

Cross-section data on enlistments and other variables in a peacetime period obviously allow one to avoid this problem. However, military pay does not vary across areas, so one must assume that the ratio of military to civilian pay (which does vary geographically) determines enlistments if one is to estimate the effect of military pay. Moreover, unmeasured differences in tastes and test scores may be correlated with civilian compensation and/or unemployment rates. For example, the historically greater propensity to enlist in the South may be due to lower civilian wage opportunities or to differences in "cultural" attitudes toward the military; failure to control for these taste differences biases the coefficients of other factors correlated with them, such as civilian pay.

When one considers the strengths and weaknesses of the time-series and cross-section

\*Department of Economics, University of Maryland, College Park, MD 20742. I have benefited from help with data sources and/or comments on a previous draft from George Abrahams, Bernard Bell, Ed Byerly, Michael Cimini, Charles Dale, Curtis Gilroy, Sandra Grove, Daniel Hamermesh, Philip Knorr, Walter Oi, Thomas Schneider, Gary Solon, Marvin Trautwein, and seminar participants at Cornell, George Washington, Maryland, Michigan State, National Bureau of Economic Research, and Virginia. Cheryl Hansen provided careful research assistance. Any remaining inadequacies are my doing. This study was financed by the U.S. Army Research Institute; additional computer support was provided by the University of Maryland Computer Science Center. The judgments expressed are mine, and should not be interpreted as reflecting the position of NBER, ARI, or the Department of Defense.

<sup>1</sup>Previous studies are discussed in more detail in my 1984 paper.



tion approaches, pooling cross-section data over the all-volunteer period emerges as an alternative worth considering. The cross-sectional dimension allows us to avoid worrying about holding constant the effects of the draft and the Vietnam war, and gives a good deal of variation in unemployment rates and some variation in civilian earnings. Pooling several cross sections allows one to introduce state-specific dummy variables to deal with state-specific differences in tastes and ability.<sup>2</sup> It also provides *some* variation in military compensation, though obviously less than would be available in longer time-series. With individual-specific intercepts, variations around state means for each of the variables over the sample period identify the coefficients, and so a short-run response is being captured.

### I. Data

The dependent variables used in this study are ratios of the number of contracts signed by male nonprior-service Army enlistees to the enlistment-age population.<sup>3</sup> Contracts cross tabulated by high school graduation and mental-test category by state for fiscal years 1976–82 were made available by the Defense Manpower Data Center. Categories (CAT) I-IIIa correspond to the top half of the test-score distribution. Four groups of recruits were analyzed: CAT I-IIIa high school graduates, all CAT I-IIIa's, all high school graduates, and all enlistees. For the first and third groups, the dependent variable is the ratio of contracts signed to number of new high school graduates in the last three years; for the other groups, it is the ratio of contracts signed to population 18–20 years of age.

Military pay is measured by basic military compensation (*BMC*), which includes the

value of allowances and tax advantages as well as base pay.

Educational benefits changed several times during the sample period. For various reasons, including the fact that such benefits are received in the future and the possibility that the benefits will not be used at all, these benefits are thought to be worth less to the recruit than their stated value. Daniel Huck, Lorin Kusmin, and Edwin Shepard (1982) calculated a range of estimates of the value of each type of educational benefit to the recruit. Using the average of the high and low estimate, the various educational benefits were combined into a single variable *ED*. If total compensation equals *BMC* plus *ED* then<sup>4</sup>

$$\begin{aligned}\ln(\text{total compensation}) \\ &= \ln(BMC) + \ln(1 + ED/BMC) \\ &= \ln(BMC) + ED/BMC.\end{aligned}$$

Military compensation consists of a large number of benefits and bonuses not included in *BMC*. Indeed, a recent study (Office of the Secretary of Defense, 1982) listed 60 categories of compensation apart from those included in *BMC*. Potential biases from ignoring these benefits can be avoided if one is willing to accept the assumption that the *ratio* of military pay to civilian pay determines enlistments. In this case, fiscal-year dummy variables can control for variations in military pay ( $W_m$ ), and the effect of both military and civilian pay determined from the coefficient of civilian pay ( $W_c$ ).

Civilian earnings  $W_c$  are measured by the quarterly average (by state) of monthly earnings of private workers, based on Unemployment Insurance (UI) records, that include nearly all private employment. The reported earnings are total earnings, not just the portion of earnings subject to UI taxes. Military and civilian pay were deflated by the Consumer Price Index.<sup>5</sup>

<sup>2</sup>Lawrence Goldberg and Peter Greenston (1983) pool time-series and cross-section data, but do not adopt the dummy-variable strategy. Instead they include the proportion black and proportion urban as controls for tastes and abilities.

<sup>3</sup>The data and data sources are described in more detail in an appendix, available from the author upon request.

<sup>4</sup>Since the *discounted ED* averaged only about 4 percent of *BMC*, the approximation  $\ln(1+x) \approx x$  is quite accurate.

<sup>5</sup>Civilian earnings were also corrected for between-state differences in living costs, using data from the

The unemployment rates used in this paper are quarterly averages of monthly unemployment rates tabulated from the *Current Population Survey (CPS)*. These were provided by the Bureau of Labor Statistics, though they emphasize that the monthly *CPS* tabulations are "official" unemployment rate estimates only for the 10 largest states. Even these unofficial tabulations were not available for about twenty of the smallest states prior to January 1976; the missing values were imputed from regressions (using 1976–82 data) which related the state unemployment rate to the national rate, the state and national employment-population ratio, and a time trend.

During this sample period, unemployment trends varied widely across regions. The unemployment rate in the Midwest as a fraction of the national average rate increased by 25 percentage points from fiscal 1976 to fiscal 1982, while those of the Northeast and West declined by 27 and 11 points, respectively. Unemployment rates in the South remained at about 90 percent of the national average throughout the period.

## II. Results

Regression results are presented in Table 1. For each dependent variable, military compensation is measured by *BMC* or a set of dummy variables. Both *OLS* and *GLS* estimates are presented.<sup>6</sup> The *GLS* estimates (that correct for first-order serial correlation) probably deserve greater weight, but the fact that the serial correlation coefficient must be estimated makes the issue less clear cut.<sup>7</sup>

---

Bureau of Labor Statistics low-income family budget series, though this correction had little impact on the estimates.

<sup>6</sup>The first-order serial correlation was estimated by the method described by Stephen Nickell (1981) and Gary Solon (1983). The equation was then quasi-first-differenced using this estimate.

<sup>7</sup>*OLS* misestimates the standard errors of coefficients in the presence of serial correlation, but in these data the standard errors changed little when *GLS* was used. *GLS* also provides smaller true standard errors when the serial correlation coefficient is known, but this gain does not necessarily occur when the correlation is estimated. In unpublished Monte Carlo experiments, Solon finds virtually no gain when using 50 cross-sectional units and 10 observations per unit when the true correlation is 0.4.

The dependent variables, *BMC*, and  $W_c$  are all entered logarithmically, so that the coefficients of *BMC* and  $W_c$  are elasticities. Given the variation in unemployment rates noted above, there seemed reason to hope that the unemployment response could be estimated with considerable precision; hence a quadratic specification was adopted. The unemployment rate is shown in percent, and coefficients of the quadratic unemployment term have been multiplied by 100. The remaining columns of the table are the coefficient of the time trend (that runs from 1 to 28), the standard error of the equation, and the elasticity of the recruiting measure with respect to the unemployment rate at the sample-mean unemployment rate of 7 percent. Not shown in the table are dummy variables for individual states,<sup>8</sup> and quarter of the year. Each state's observations were weighted by the average number of persons 18–20-years old in the state.

The dependent variable in the first four equations of Table 1 is the logarithm of the number of CAT I-III high school graduate contracts divided by the number of high school graduates. The estimated elasticity of supply with respect to military compensation is about 0.5 when *BMC* is used, but the implied elasticity is 1.0 to 1.5 when the dummy-variable approach is used. Civilian wages are negatively related to enlistments. In the equations with *BMC*, the restriction that supply is determined by the ratio of military to civilian pay (i.e., that coefficients of *BMC* and  $W_c$  be equal and opposite-signed) cannot be rejected.

The unemployment rate has very large effects on enlistment. At the mean unemployment rate of 7 percent, a one-percentage-point reduction in the unemployment rate reduces enlistments by about 10 percent. Given the difficulty of estimating unemployment effects with any precision with time-series data (see Ash, Udis, and McNown), the significance of these estimates is worth underlining.

<sup>8</sup>To reduce computational costs, the equations were actually estimated by deviating all variables from their state-specific means, which is equivalent to including state-specific dummy variables.

TABLE 1—ENLISTMENT EQUATIONS

	<i>BMC</i>	<i>Ed/BMC</i>	<i>W<sub>c</sub></i>	<i>UR</i>	<i>UR</i> <sup>2</sup>	Trend	<i>S. E.</i>	<i>p</i>	<i>n<sub>U</sub></i>
<b>A</b>	.61	10.8	-1.04	.17	-.34	.019	.2358		.86
	(2.5)	(38.2)	(3.8)	(9.7)	(3.3)	(13.3)			
	.44	9.9	-.17	.10	-.05	.028	.2057	.42	.68
	(1.5)	(30.0)	(.6)	(5.5)	(.5)	(15.2)			
	<sup>a</sup>	<sup>a</sup>	-1.00	.10	-.06	<sup>a</sup>	.2468		.65
<b>B</b>			(3.4)	(5.1)	(.5)				
	<sup>a</sup>	<sup>a</sup>	-1.50	.09	.01	<sup>a</sup>	.2382	.11	.64
			(6.1)	(4.5)	(.1)				
	.96	9.1	-1.52	.11	-.15	.002	.2074		.60
	(4.5)	(36.5)	(6.3)	(6.9)	(1.7)	(1.6)			
<b>C</b>	.95	8.3	-.99	.06	.06	.009	.1827	.34	.49
	(3.8)	(30.0)	(3.9)	(3.7)	(.6)	(6.3)			
	<sup>a</sup>	<sup>a</sup>	-1.59	.07	-.02	<sup>a</sup>	.2320		.49
			(5.9)	(3.9)	(.2)				
	<sup>a</sup>	<sup>a</sup>	-.09	.06	.03	<sup>a</sup>	.2268	.07	.44
<b>D</b>			(.4)	(3.1)	(.3)				
	-1.28	6.9	-.55	.10	-.09	.007	.2103		.65
	(5.9)	(27.5)	(2.2)	(6.6)	(.9)	(5.3)			
	-2.06	7.3	.22	.07	.03	.011	.1874	.50	.53
	(7.1)	(23.1)	(.8)	(4.1)	(.3)	(5.9)			
<b>E</b>	<sup>a</sup>	<sup>a</sup>	-1.14	.08	-.01	<sup>a</sup>	.2130		.55
			(4.6)	(4.7)	(.1)				
	<sup>a</sup>	<sup>a</sup>	-1.21	.08	.01	<sup>a</sup>	.2046	.26	.55
			(5.3)	(4.1)	(.1)				
	-3.06	4.8	-.75	.02	.21	-.025	.2019		.32
<b>F</b>	(14.6)	(19.7)	(3.2)	(1.1)	(2.4)	(20.8)			
	-4.07	5.5	-.59	.02	.22	-.025	.1808	.51	.33
	(14.4)	(17.9)	(2.2)	(.9)	(2.2)	(14.0)			
	<sup>a</sup>	<sup>a</sup>	-1.85	.05	.02	<sup>a</sup>	.2055		.36
			(7.7)	(3.0)	(.2)				
<b>G</b>	<sup>a</sup>	<sup>a</sup>	1.00	.04	.01	<sup>a</sup>	.2196	.26	.32
			(4.1)	(2.2)	(.1)				

Note: Sample: 51 states, quarterly, from 1975:4 to 1982:3 = 1428 observations. The dependent variables are **A** = CAT I-III high school graduate contracts/high school graduates; **B** = CAT I-III contracts/population; **C** = high school graduate contracts/high school graduates; **D** = contracts/population. The *t*-statistics are in parentheses below coefficients. Each coefficient of the quadratic unemployment term is multiplied by 100.

<sup>a</sup> = *BMC*, *ED/BMC*, and *TREND* replaced by fiscal-year dummies.

The most anomalous result is the very large coefficient of *ED/BMC*. In theory, its coefficient should equal that of *BMC*; in fact, it is much larger. Part of the explanation may be that the procedure used in deriving this variable overdiscounts the value of the benefits, but it is hard to argue that this is a full explanation. In any case, since *ED/BMC* was highest when unemployment was highest in the sample period, overestimating the effect of *ED/BMC* is likely to lead to underestimating the unemployment effects.

The second set of equations focuses on the supply of CAT I-III enlistments (both

graduates and nongraduates). The conclusions that emerge are very similar to those reached for CAT I-III graduates. Compensation elasticities tend to be a bit larger, and the unemployment elasticities are a bit smaller. The unemployment elasticities remain important in practical terms.

When supply is measured by the proportion of high school graduates who enlist (third set of equations), the results are less satisfactory. Taken at face value, the first two specifications suggest that educational benefits significantly increase recruitment, but increased *BMC* reduces it. The dummy-variable specifications show quite clearly,

however, that recruitment success is inversely related to civilian alternatives, and the estimated elasticity ( $-1.1$  or  $-1.2$ ) is fairly sizeable. The effect of military compensation is just not reliably estimated in this set of equations.

Estimated unemployment elasticities, however, remain significant in both a statistical and practical sense. Moreover, they are not very sensitive to the choice of military pay specification.

The difficulties of estimating the effect of compensation on the ratio of total contracts to population is similar to that encountered in analyzing the high school graduate measure. Neither of the estimates based on explicit measures of military compensation (first two lines of this set) is plausible. The compensation elasticity of 1.85 implied by the *OLS* dummy variable specification is right-signed, but *GLS* estimation reverses this.

Although the estimated effects of unemployment on total contracts are smaller than for the other dependent variables, they are once again statistically significant and consistent across the three specifications. Given the relatively large percentage changes in unemployment rates that have occurred in the past decade, an elasticity of one-third implies nontrivial percentage changes in the number of recruits.<sup>9</sup>

<sup>9</sup>For a slightly shorter sample period (1976:4 through 1982:2), measures of recruitment effort could be included as explanatory variables. The effects of military compensation, civilian earnings, and unemployment are broadly similar to those in Table 1. The most consistent difference is a slightly smaller unemployment effect. For the three high-quality enlistee groups, the recruitment effort variables tell a generally sensible story. Additional Army recruiters increase enlistments, but increases in other services' recruiters reduce them. Proportional increases in all services' recruiters increase army enlistments. National media advertising sometimes has a positive effect, but this is unstable across specifications. Local media advertising seems to have no detectable positive effect on recruitment. One interpretation is that such advertising is concentrated on areas where enlistments are below expectations, but the limited available evidence seems inconsistent with this explanation (Morey and McCann). The estimates for all enlistees do not show a negative effect of DOD recruiters on army enlistments, but do show consistent positive national advertising coefficients.

An unavoidable question is why the compensation elasticities are more satisfactory for the first two supply measures than for the last two. While the difficulties of measuring such compensation are certainly important, the rough similarity among the alternative estimates for each of the first two measures suggest that this is an unlikely explanation for why the last two measures do not exhibit such stability. A more plausible explanation is that the number of CAT I-III A enlistees is supply determined, but the number of total enlistees is largely demand determined. The simplest story along these lines—that changes in standards offset any changes in the supply of enlistees—would predict that estimated effects would be smaller than true ones, perhaps even zero, but would not lead one to expect wrong-signed “effects.” This would seem to require that the total number of recruits demanded be negatively related to the offered wage. While military manpower demands are not usually thought of as having a significant demand elasticity in the short run, such a negative correlation could arise either by chance (recall there are only seven fiscal years of data), or because higher military compensation reduces the demand for *new* recruits by improving retention. If, in a period of excess supply, recruitment standards are raised *uniformly*, one might expect that those variables with significant geographic variation would determine how the demand-limited total of enlistments would be distributed across the country. Thus, demand constraints might be consistent not only with wrong-signed estimates for military pay, but right-signed estimates of the impacts of civilian pay and unemployment. While Arthur De Vany and Thomas Saving (1982) have modeled the way in which discrepancies between the supply of enlistees of the desired quality and the desired number of such enlistees is reconciled at the national level, there seems to be little available research on how demand constraints make themselves felt locally.

### III. Conclusions

In order to achieve an adequate sample size without including draft-period observa-

tions, and to take advantage of large regional differences in the path of unemployment in recent years, the determinants of army enlistments were estimated from quarterly data by state for fiscal years 1976–82. In sharp contrast to several previous papers, unemployment rates had quite strong effects on recruitment success. For various categories of high-quality recruits, the elasticity of contracts signed with respect to the unemployment rate ran from .4 to .8. For high-quality recruits, defined as those with scores in the top half of the distribution on the military's entrance test (or those with these scores and high school degrees), estimates of the elasticity of contracts with respect to military compensation centered *roughly* on 1.0. For total contracts and contracts signed by high school graduates (regardless of test scores), the compensation elasticity could not be estimated with any confidence. This may be because the number of such enlistments is demand constrained rather than supply determined. There was, however, consistent evidence that the number of contracts was inversely related to alternative (civilian) earnings.

The combination of sizeable unemployment elasticities and major swings in the regional concentration of unemployment had significant effects on the regional distribution of Army enlistees. During the sample period, *the Midwest replaced the South as the dominant per capita supplier of recruits to the volunteer Army.*

These results suggest that, as unemployment rates fall, the Army will have difficulty maintaining the number of high-quality recruits at currently offered compensation levels. Indeed, this adjustment has already begun: from the first half of fiscal 1983 to the first half of fiscal 1984, the unemployment rate declined from 10 to 8 percent, and unpublished U.S. Army Recruiting Command data show an 18 percent decline in the number of CAT I-III high school graduate recruits.

There are several options for dealing with this difficulty within the all-volunteer framework. The most obvious is to increase the compensation of new enlistees. A subtler strategy is to focus on retention of those

already in the Army, reducing the number of high-quality enlistees who must be recruited.<sup>10</sup> A third option is to expand the role of women. A final issue is whether the recent numbers of high-quality recruits *ought* to be maintained, since those levels were caused by the recession. This involves comparing the marginal value of relatively able individuals in military and civilian employment.

Two quite different directions for future research seem desirable, given the results of this paper. First, research on how recruiting standards respond to local or national shortages or surpluses of recruits meeting a given standard would be very useful. Second, the lag (if any) in the response of enlistments to military and civilian compensation, unemployment, etc., deserves greater attention.

<sup>10</sup> Educational benefits may be effective in encouraging new high-quality enlistments, since they are a way of paying college-eligible recruits more than others. However, because they are valuable only when one returns to civilian life, their effect on retention may be perverse. Allowing those who have earned educational benefits but remain in the military to transfer their benefits to spouses or children has been proposed as a way of blunting the disincentive to reenlist.

## REFERENCES

- Ash, Colin, Udis, Bernard and McNown, Robert F., "Enlistments in the All-Volunteer Force: A Military Personnel Supply Model and Its Forecasts," *American Economic Review*, March 1983, 73, 145–55.
- Brown, Charles, "Military Enlistments: What Can We Learn from Geographic Variation?," Working Paper 1261, National Bureau of Economic Research, January 1984.
- Dale, Charles and Gilroy, Curtis, "The Effect of the Business Cycle on the Size and Composition of the U.S. Army," *Atlantic Economic Journal*, March 1983, 11, 42–53.
- De Vany, Arthur S. and Saving, Thomas R., "Life Cycle Job Choice and the Demand and Supply of Entry-Level Jobs—Some Evidence from the Air Force," *Review of Economics and Statistics*, August 1982, 64, 457–65.
- Goldberg, Lawrence and Greenston, Peter, "A

Time-Series, Cross-Sectional Study of Enlistment Supply," paper presented at meeting of the Southern Economic Association, 1983.

Huck, Daniel, Kusmin, Lorin and Shepard, Edwin, "Improving Educational Benefits: Effects on Costs, Recruiting, and Retention," U.S. Congressional Budget Office, March 1982.

Morey, Richard C. and McCann, John M., "Armed Services Recruiting Research: Issues, Findings, and Needs," unpublished

paper, 1983.

Nickell, Stephen, "Biases in Dynamic Models with Fixed Effects," *Econometrica*, November 1981, 49, 1417-26.

Solon, Gary, "Estimating Autocorrelations in Fixed Effects Models," Working Paper No. 160, Industrial Relations Section, Princeton University, April 1983.

Office of the Secretary of Defense, *Military Compensation Background Papers*, 2nd ed., Washington: USGPO, 1982.

# Paying for Public Inputs

By RICHARD MANNING, JAMES R. MARKUSEN, AND JOHN McMILLAN\*

With public consumption goods, Pareto optimality can be achieved in equilibrium through Lindahl pricing. This requires that each consumer pays a price proportional to his marginal utility from the public good. The marginal cost of providing the good is then equated to the sum of the marginal benefits.

Some public goods serve as inputs into production processes rather than as consumption goods: the lighthouse is an example. Efficiency in an economy with public intermediate goods requires that the marginal cost of providing them equals the sum of their marginal benefits to firms. By analogy with the case of public consumption goods, it might seem that Lindahl pricing can be extended to public intermediate goods, by requiring that firms pay in proportion to the marginal contribution of these goods to profits. Agnar Sandmo (1972) explicitly proposes this for convex technologies. For nonconvex technologies that arise with an important class of production functions (essentially those exhibiting "atmosphere externalities," in the phrase of James Meade, 1952), Lindahl pricing for such public inputs is infeasible.

The principal result derived here is a simple alternative to Lindahl pricing of public inputs. This is an across-the-board tax on factor incomes at a rate that depends on production parameters. Efficiency is guaranteed if this tax is used.<sup>1</sup>

Means of paying for public inputs are discussed after the economic structure is defined in the next section. Some comments on the analysis are provided in the concluding section.

## I. The Economic Structure

Private consumption goods are produced according to

$$(1) \quad Y_i = F_i(X_i, R), \quad i = 1, \dots, n,$$

where  $Y_i$  is the output of,  $X_i \equiv (X_{i1}, \dots, X_{im})$  is a vector of private-factor inputs to, and  $F_i$  is the production function of, industry  $i$ . The variable  $R$  is the amount of public intermediate good produced and available to all firms. This good is produced from primary factors alone,

$$(2) \quad R = G(X_{n+1}),$$

where  $X_{n+1} \equiv (X_{n+1,1}, \dots, X_{n+1,m})$ , and  $G$  is the production function for public intermediate goods.

For simplicity,  $F_i$ ,  $i = 1, \dots, n$ , and  $G$  are assumed to be continuously differentiable, and  $F_{ik}$  denotes  $\partial F_i / \partial X_{ik}$ , etc. In addition,  $G$  is taken to be quasi concave and linearly homogeneous, and  $F_i$ ,  $i = 1, \dots, n$  is concave with respect to private inputs. This is, of course, the usual assumption of neoclassical general equilibrium theory. Given  $X_{n+1}$ , matters are as in that theory, therefore.

Every primary factor  $k$  is in fixed supply  $\bar{X}_k$ , so that

$$(3) \quad \sum_{i=1}^{n+1} X_{ik} = \bar{X}_k, \quad k = 1, \dots, m.$$

Efficiency in the economy described by (1), (2), and (3) requires that

$$(4) \quad r_k = p_i F_{ik}, \quad i = 1, \dots, n; \quad k = 1, \dots, m,$$

where  $r_k$  is the return to the  $k$ th private input, and  $p_i$  is the price of the  $i$ th consumption good. In addition, the Samuelson condition

$$(5) \quad \sum_{i=1}^n \frac{F_{ir}}{F_{ik}} = \frac{1}{G_k}, \quad k = 1, \dots, m$$

\*Departments of Economics; Manning: University of Canterbury, Christchurch, New Zealand; Markusen and McMillan: University of Western Ontario, London N6A 5C2 Canada.

<sup>1</sup>It is assumed that because of free entry—the public input is available for anyone to use—the industry benefiting from the public input is perfectly competitive.

must be satisfied. The efficiency conditions (4) and (5) are essentially those reached by Keimei Kaizuka (1965), Takashi Negishi (1973), and Sandmo.

With public intermediate goods there is a presumption that the production functions of the consumption-goods industries exhibit increasing returns to scale. For the case of one primary factor (called "labor"), this may be argued as follows:

With private inputs the assumption of constant returns to scale is traditionally justified by the replication of basic processes. With public inputs the concept of replication applies in the following way to production functions: doubling the number of workers in an industry, keeping the supply of the public factor fixed, leaves each worker with the same amount of the public factor to work with, so that, by the hypothesis of replication, output will double.

[Manning-McMillan, 1979, p. 246]

When there are many primary factors the argument is the same. Thus, replication implies that  $F_i$ ,  $i = 1, \dots, n$  is linearly homogeneous in the private inputs; that is,

$$(6) \quad F_i(\lambda X_i, R) = \lambda F_i(X_i, R),$$

$$\lambda > 0, \quad i = 1, \dots, n.$$

This form of production function was also used by M. Ali Khan (1980) and Thomas Pugel (1982). Clearly, (6) implies increasing returns to scale to all inputs. Means of paying for public inputs in competitive market economies when (6) applies have not been adequately discussed, so this issue is now taken up.

## II. Means of Payment for Public Inputs

Following Sandmo, the *Lindahl prices* of the public intermediate good are defined by

$$(7) \quad \pi_i = p_i F_{ir}, \quad i = 1, \dots, n.$$

Such prices generate a Pareto optimum if, for example, the production functions for

consumption goods exhibit constant returns to scale with respect to all inputs, as in Negishi and Makoto Tawada (1980). Consider, however, the problem faced by a competitive firm when (6) applies. The firm faces parametrically given prices for private goods and factors, and a given supply of the public input. It must solve

$$(8) \quad \max_{X_i \geq 0} p_i F_i(X_i, R) - \sum_{j=1}^m r_j X_{ij} - \pi_i R.$$

The following is obvious, although a proof is in order.

*Remark:* When production processes permit replication, Lindahl pricing in a competitive market economy implies that the equilibrium output of firms is zero.

### PROOF:

Suppose that  $X_i \geq 0$ .<sup>2</sup> Then (8) implies that (4) is satisfied. However, Euler's Theorem implies that

$$(9) \quad F_i(X_i, R) = \sum_{k=1}^m F_{ik} X_{ik}$$

in view of (6). Thus, profits are negative, since

$$(10) \quad p_i F_i(X_i, R) - \sum_{j=1}^m r_j X_{ij} - \pi_i R = -\pi_i R < 0.$$

But now suppose that  $X_i = 0$ . Then (6) implies

$$(11) \quad F_i(0, R) = 0.$$

This implies that

$$(12) \quad F_{ir}(0, R) = 0,$$

so that profits are zero if no inputs are used. That is

$$(13) \quad p_i F_i(0, R) - \pi_i R = 0.$$

<sup>2</sup>That is, not all private inputs are zero.



Together, (10), (11), and (13) imply that the firm has no output because it maximizes profit by using no private inputs.

This demonstrates the infeasibility of Lindahl pricing for public inputs in a competitive market economy when the technology is of a reasonable type. Notice the importance of replication in ensuring (12), and hence (13). It will now be shown that there is a simple formula for the proportional income tax rate which generates enough revenue to pay for an efficient quantity of the public input. This formula will involve the parameters

$$(14) \quad \eta_i \equiv F_{ir}R/Y_i, \quad i=1, \dots, n,$$

where  $0 \leq \eta_i \leq 1$  is the elasticity of the  $i$ th consumption-good output with respect to public good supply.<sup>3</sup> Define

$$(15) \quad \alpha_i = p_i Y_i / \sum_{i=1}^n p_i Y_i, \quad i=1, \dots, n,$$

to be the  $i$ th industry's share of the value of output of consumption goods. Further, define

$$(16) \quad \bar{\eta} \equiv \sum_{i=1}^n \eta_i \alpha_i.$$

**PROPOSITION:** *In a competitive market economy, the optimal amount of a public input is supplied if income taxes collected at the rate  $T = \bar{\eta}/(1 + \bar{\eta})$  are all spent on the public input.*

**PROOF:**

From (4) and (5) it follows that

$$(17) \quad r_j = p_i F_{ij} = \sum_{i=1}^n p_i F_{ir} G_j.$$

Thus

$$(18) \quad r_j/r_k = \sum_{i=1}^n p_i F_{ir} G_j / p_i F_{ik}.$$

Multiply through (18) by  $r_k X_{n+1,j}$  and multiply and divide the right-hand side by  $R$ . This gives

$$(19) \quad r_j X_{n+1,j} = \frac{\left( \sum_{i=1}^n p_i F_{ir} R \right) (X_{n+1,j} G_j / R)}{(p_i F_{ik} / r_k)}.$$

Notice that the definition (14) implies that

$$(20) \quad \sum_{i=1}^n p_i F_{ir} R = \sum_{i=1}^n p_i Y_i \left( \frac{F_{ir} R}{Y_i} \right) = \sum_{i=1}^n p_i Y_i \eta_i.$$

The linear homogeneity of  $G$  implies that

$$(21) \quad \sum_{j=1}^m X_{n+1,j} G_j = R.$$

Summing (19) over  $j=1, \dots, m$ , and using (4), (20), and (21) implies that<sup>4</sup>

$$(22) \quad \sum_{j=1}^m r_j X_{n+1,j} = \sum_{i=1}^n p_i Y_i \eta_i.$$

The left-hand side of (22) is the expenditure needed to produce the public input. Total income (including rents, if any, due to decreasing returns to private factors) is just

$$(23) \quad \sum_{j=1}^m r_j X_{n+1,j} + \sum_{i=1}^n p_i Y_i.$$

The assumptions of a balanced budget and a proportional income tax imply, from (22)

<sup>3</sup>Manning and McMillan introduced these parameters in their earlier paper, and showed their importance in determining the shape of production possibilities. See also Tawada and K. Abe (1984).

<sup>4</sup>If there are nonconstant returns to scale in producing public inputs, of course (21) does not hold. However, (22) can be modified by including at the right-hand side a term reflecting the departure from constant returns. This will also appear in the tax formula (24).

and (23), that

$$(24) \quad T = \frac{\sum_{i=1}^n p_i Y_i \eta_i}{\sum_{i=1}^n p_i Y_i \eta_i + \sum_{i=1}^n p_i Y_i}.$$

The definition (15) gives the tax formula from (24), on division through numerator and denominator by  $\sum_{i=1}^n p_i Y_i$ .

### III. Concluding Remarks

It is frequently argued that efficiency requires that public services be financed by direct charges on their users. It is shown here to be sometimes impossible to apply user charges. However, the proposition suggests a procedure which will work even whether or not user charges are feasible. Furthermore, the proportional tax system is administratively much more straightforward than Lindahl pricing, so it is to be preferred on that criterion as well.

While the economic structure considered here is much simplified,<sup>5</sup> it is interesting that the parameters determining the efficient tax rate are, in principle, capable of being estimated. This suggests that the conclusions reached here are of some relevance in policymaking. In this connection, it might be noted that an ad valorem sales tax, imposed on sales of the private good at the same rate as the efficient income tax, is equally good.

The problem solved in this note is similar to some found elsewhere. For instance, John Chipman (1970) provides an ideal tax-subsidy scheme for an economy with external economies of scale. In his scheme an industry is taxed (subsidized) if its elasticity of output with respect to scale is less (greater)

than the weighted average elasticity of all industries, with expenditure shares being the weights. The same, in effect, occurs in the present case. While everyone is taxed at the same rate, industries that are above average in sensitivity to public inputs gain most, so they get a net subsidy.

### REFERENCES

- Chipman, John S., "External Economies of Scale and Competitive Equilibrium," *Quarterly Journal of Economics*, August 1970, 84, 347-85.
- Kaizuka, Keimei, "Public Goods and Decentralization of Production," *Review of Economics and Statistics*, February 1965, 47, 118-20.
- Khan, M. Ali, "A Factor Price and Public Input Equalization Theorem," *Economics Letters*, 1980, 5, 1-6.
- Manning, R. and McMillan, J., "Public Intermediate Goods, Production Possibilities, and International Trade," *Canadian Journal of Economics*, May 1979, 12, 243-57.
- Meade, James E., "External Economies and Diseconomies in a Competitive Situation," *Economic Journal*, March 1952, 62, 54-67.
- Negishi, Takashi, "The Excess of Public Expenditures on Industries," *Journal of Public Economics*, July 1973, 2, 231-40.
- Pugel, Thomas A., "Endogenous Technological Change and International Technology Transfer in a Ricardian Trade Model," *Journal of International Economics*, November 1982, 13, 321-35.
- Sandmo, Agnar, "Optimality Rules for the Provision of Collective Factors of Production," *Journal of Public Economics*, April 1972, 1, 149-57.
- Tawada, Makoto, "The Production Possibility Set with Public Intermediate Goods," *Econometrica*, May 1980, 48, 1005-12.
- \_\_\_\_\_ and Abe, K., "Production Possibilities and International Trade with a Public Intermediate Good," *Canadian Journal of Economics*, May 1984, 17, 232-48.

<sup>5</sup>Note that the results of this analysis rely on the inelasticity of factor supplies.

# Excess Labor and the Business Cycle

By RAY C. FAIR\*

In a remarkable empirical study of 168 U.S. manufacturing plants, James Medoff and Jon Fay (1985) have examined the magnitude of labor hoarding during economic contractions. They found that during its most recent trough quarter, the typical plant paid for about 8 percent more blue-collar hours than were needed for regular production work. Some of these hours were used for other worthwhile work, and after taking account of this, 5 percent of the blue-collar hours was estimated to be hoarded for the typical plant.

The hypothesis that firms may hold "excess labor" during contractions was explored in my 1969 study, using monthly three-digit industry data. A model of labor demand was developed that is based on the idea that firms may at times hold excess labor. This model was originally estimated using the monthly three-digit industry data, and it was later estimated using aggregate quarterly data. The aggregate labor demand equations are part of my U.S. macro model. The latest discussion of the aggregate equations is in chapter 4 in my 1984 study. Both the monthly industry estimates and the quarterly macro estimates support the excess labor hypothesis.

The purpose of this paper is to see if the quantitative estimates of Medoff and Fay are consistent with the aggregate estimates. If this is the case, which the results in this paper show, it provides a strong argument in favor of the excess labor hypothesis. Essentially the same conclusion has been reached using two very different data sets. This is one of the few examples in macroeconomics where a hypothesis has been so strongly confirmed using detailed micro data.

## I. Review of the Aggregate Labor Demand Equations

The latest discussion of the theoretical model upon which the labor demand equations are based is in chapter 3 of my 1984 study. Only a few features of this model will be reviewed here. The technology is assumed to be putty-clay, where at any one time there are a number of different types of machines that can be purchased. The machines differ in price, in the number of workers that must be used with each machine per unit of time, and in the amount of output that can be produced per machine per unit of time. The worker-machine ratio is assumed to be fixed for each type of machine. Adjustment costs are postulated for changes in the size of the work force and for changes in the size of the capital stock. Firms behave by maximizing the present discounted value of expected future after-tax cash flow. The main decision variables of a firm are its price, production, investment, labor demand, and wage rate. Because of the adjustment costs, it may sometimes be optimal for a firm to operate "off" its production function and hold excess labor and/or excess capital.

The transition from a theoretical to an econometric model is always difficult in macroeconomics, and the present case is no exception. This transition is discussed in chapter 4 of my 1984 study, and again only a few features will be discussed here. For the empirical work the production function is postulated to be one of fixed proportions:

$$(1) \quad Y = \min\{\lambda(J \cdot H^J), \mu(K \cdot H^K)\},$$

where  $Y$  is production,  $J$  is the number of workers employed,  $H^J$  is the number of hours worked per worker,  $K$  is the stock of capital,  $H^K$  is the number of hours each unit of  $K$  is utilized, and  $\lambda$  and  $\mu$  are coefficients that may change over time due to technical

\*Department of Economics, Yale University, New Haven, CT 06520. I am indebted to a referee for helpful comments.

progress. The variables  $Y$ ,  $J$ , and  $K$  are observed;  $H^J$  and  $H^K$  are not. This production function is only an approximation to the technology of the theoretical model. It does not allow for the existence of more than one type of machine, and it treats technical progress in an inappropriate way. Even if there were only one type of machine in existence, technical progress would take the form of machines having different  $\lambda$  and  $\mu$  coefficients depending on when they were purchased. In order to account for technical progress in this way, one would have to keep track of when each machine was purchased and what the coefficients were for that machine. This kind of detail is not possible with aggregate data, and one must resort to simpler specifications.

Given the production function, the next step is to measure the number of worker hours required to produce the output each period. This was done as follows. Output per paid-for worker hour,  $Y/(J \cdot H)$ , was first plotted for the 1952:I–1982:III period. (Data on hours paid for,  $H$ , exist, whereas data on hours worked,  $H^J$ , do not.) The peaks of this series were assumed to correspond to cases where the number of hours worked equals the number of hours paid for (i.e., where  $H^J = H$ ), which implies that values of  $\lambda$  in equation (1) are observed at the peaks. The values of  $\lambda$  other than those at the peaks were then assumed to lie on straight lines between the peaks. Given an estimate of  $\lambda$  for a particular quarter and given the production function (1), the estimate of the number of worker hours required to produce the output of the quarter (denoted  $JHMIN$ ) is simply  $Y/\lambda$ . The peaks that were used for the interpolations are 1952:I, 1953:II, 1955:I, 1966:I, 1973:I, and 1977:I. The line connecting 1973:I and 1977:I was extrapolated beyond 1977:I to fill out the series through 1982:III.

In the theoretical model, a firm's price, production, investment, labor demand, and wage rate decisions are made simultaneously in the sense that all are derived from the solution of the firm's maximization problem. For the empirical work the decisions are assumed to be made sequentially, where the sequence is price, production, investment,

labor demand, and wage rate. The labor demand equations are thus based on the assumption that the production decision has already been made. Were it not for the adjustment costs of changing employment, the optimal level of employment would merely be the amount needed to produce the output of the period, but, because of these costs, excess labor may be held during certain periods. In the theoretical model there was no need to postulate explicitly how employment deviates from the amount required to produce the output, but this must be done for the empirical work.

The estimated demand-for-workers equation is based on the following three equations:

$$(2) \quad \Delta \log J = \alpha_0 \log \frac{J_{-1}}{J_{-1}^*} + \alpha_1 \Delta \log Y \\ + \alpha_2 \Delta \log Y_{-1} + \alpha_3 \Delta \log Y_{-2},$$

$$(3) \quad J_{-1}^* = JHMIN_{-1}/H_{-1}^*,$$

$$(4) \quad H_{-1}^* = \bar{H}e^{\delta t},$$

where  $JHMIN$  is the number of worker hours required to produce the output of the period,  $H^*$  is the average number of hours per worker that the firm would like to be worked if there were no adjustment costs, and  $J^*$  is the number of workers the firm would like to employ if there were no adjustment costs. The term  $\log(J_{-1}/J_{-1}^*)$  in equation (2) will be referred to as the (logarithmic) "number of excess workers" on hand. Equation (2) states that the change in the demand for workers is a function of the number of excess workers on hand and three change-in-output terms (all changes are changes in logs). If output has not changed for three periods and if there are no excess workers on hand, the change in workers employed is zero. The change-in-output terms are means in part to be proxies for expected future output changes. Equation (3) defines the desired number of workers, which is simply equal to the required number of worker hours divided by the desired number of hours worked per worker. Equation (4) postulates that the desired number of hours worked is a smoothly

trending variable, where  $\bar{H}$  and  $\delta$  are constants.

Combining equations (2)–(4) yields

$$(5) \quad \Delta \log J = \alpha_0 \log \bar{H} \\ + \alpha_0 \log \frac{J_{-1}}{JHMIN_{-1}} + \alpha_0 \delta t + \alpha_1 \Delta \log Y \\ + \alpha_2 \Delta \log Y_{-1} + \alpha_3 \Delta \log Y_{-2}.$$

This equation was estimated by two-stage least squares under the assumption of first-order serial correlation of the error term for the 1954:I–1982:III period. The estimated equation is ( $t$ -statistics in absolute value are shown in parentheses):<sup>1</sup>

$$(6) \quad \Delta \log J = -.885 - .141 \log \frac{J_{-1}}{JHMIN_{-1}} \\ (3.76) \quad (3.75) \\ + .000176t + .281 \Delta \log Y \\ (4.28) \quad (8.33) \\ + .119 \Delta \log Y_{-1} + .033 \Delta \log Y_{-2} \\ (3.03) \quad (1.02) \\ - .00967 D593 + .00174 D594 \\ (2.70) \quad (0.50)$$

$$SE = .00355, R^2 = .780, D-W = 2.04, \hat{\rho} = .447 \\ (4.44)$$

where  $D593$  and  $D594$  are dummy variables for the 1959 steel strike. The estimated value of  $\alpha_0$  is  $-.141$ , which means that, other things being equal, 14.1 percent of the number of excess workers on hand is eliminated each quarter. The implied value of  $\bar{H}$  is 531.97, which at a weekly rate is 40.92 hours. The implied value of  $\delta$  is  $-.00125$ . The trend variable  $t$  is equal to 9 for the first quarter of the sample period (1954:I), and so the implied value of  $H_{-1}^*$  for 1954:I at a weekly rate is  $40.92 \cdot \exp(-.00125 \times 9) = 40.46$ . For 1982:III,  $t$  is equal to 123, and so

the implied value for this quarter is  $40.92 \cdot \exp(-.00125 \times 123) = 35.09$ . In general these numbers seem reasonable.

The estimated demand-for-hours equation is based on equations (3), (4), and the following equation:

$$(7) \quad \Delta \log H = \lambda \log(H_{-1}/H_{-1}^*) \\ + \alpha_0 \log(J_{-1}/J_{-1}^*) + \alpha_1 \Delta \log Y.$$

The first term on the right-hand side of equation (7) is the (logarithmic) difference between the actual number of hours paid for per worker in the previous period and the desired number. The reason for the inclusion of this term in the demand-for-hours equation but not in the demand-for-workers equation is that, unlike  $J$ ,  $H$  fluctuates around a slowly trending level of hours. This restriction is captured by the first term in (7). The other two terms are the number of excess workers on hand and the current change in output. Both of these terms have an important effect on the demand-for-workers decision, and they should also affect the demand-for-hours decision since the two decisions are closely related. Past output changes might also be expected to affect the demand-for-hours decision, but these were not found to be significant and so are not included in (7).

Combining (3), (4), and (7) yields

$$(8) \quad \Delta \log H = (\alpha_0 - \lambda) \log \bar{H} \\ + \lambda \log H_{-1} + \alpha_0 \log \frac{J_{-1}}{JHMIN_{-1}} \\ + (\alpha_0 - \lambda) \delta t + \alpha_1 \Delta \log Y.$$

The estimated equation is

$$(9) \quad \Delta \log H = 1.37 - .284 \log H_{-1} \\ (4.95) \quad (5.16) \\ - .0659 \log \frac{J_{-1}}{JHMIN_{-1}} - .000250t \\ (3.55) \quad (4.94) \\ + .120 \Delta \log Y \\ (4.40)$$

$$SE = .00285, R^2 = .398, D-W = 2.18.$$

The estimated value of  $\lambda$  is  $-.284$ , which

<sup>1</sup>The first-stage regressors that were used for this work are presented in Table 6-1 in my earlier study (1984). The same holds for equation (9) below.

means that, other things being equal, actual hours per worker are adjusted towards desired hours by 28.4 percent per quarter. The excess workers variable is significant, with an estimated value of  $\alpha_0$  of  $-.0659$ . The implied value of  $\bar{H}$  is 534.60, which is 41.12 hours at a weekly rate. This compares closely to the value of 40.92 implied by equation (6). The implied value of  $\delta$  is  $-.00115$ , which compares closely to the value of  $-.00125$  implied by equation (6). No attempt was made to impose the restriction that  $\bar{H}$  and  $\delta$  are the same in equations (6) and (9). Given the closeness of the estimates, it is unlikely that imposing this restriction would make much difference.

The significance of the excess workers variable in equations (6) and (9) provides support for the excess labor hypothesis. It seems unlikely that a variable like this would be significant if firms never or seldom held excess labor.

## II. Comparison

The main concern of this paper is whether the above aggregate empirical results are consistent with the Medoff-Fay micro results. Before making this comparison, various concepts of "excess" labor need to be reviewed. Medoff and Fay distinguish between regular production work and other work. Much of the other work is maintenance. They find that at its trough in output the typical plant paid for about 8 percent more hours than were needed for regular production work. About 3 of this 8 percent was used for worthwhile other work, which means that about 5 percent of the hours was truly hoarded. Firms appear to shift at least some maintenance work from high-output to low-output periods.

For the aggregate work above there is no distinction between regular production work and other work. Within this framework there are two concepts of excess labor. One is  $J/J^*$ , which is the ratio of the actual number of workers to the long-run desired number. The other is  $(J \cdot H)/JHMIN$ , which is the ratio of total worker hours paid for to the total number required to produce the output. Note that  $J/J^*$  equals  $(J \cdot H^*)/JHMIN$ ,

where  $H^*$  is the long-run desired number of hours worked per worker.  $J/J^*$  measures how far the firm is from its long-run desired number of workers. It seems to be the appropriate "excess labor" variable to use in the labor demand equations, and it has been so used.  $(J \cdot H)/JHMIN$ , on the other hand, measures the number of hours paid for but not worked, and it seems to be the appropriate variable to compare to the Medoff-Fay estimates. It will be called the "percentage of excess hours."

If maintenance work is shifted from high- to low-output periods, then  $JHMIN$  is a misleading estimate of worker hour requirements. In a long-run sense,  $JHMIN$  is too low because it has been based on the incorrect assumption that the peak productivity values could be sustained over the entire business cycle. This error is not a serious one from the point of view of estimating the labor demand equations (6) and (9) above. If the same percentage error has been made at each peak, which is likely to be approximately the case, the error will merely be absorbed in the estimates of the constant term in the two equations. It does mean, however, that  $(J \cdot H)/JHMIN$  should not be compared to the Medoff-Fay concept of hoarded hours (i.e., to the 5 percent number). It is likely to be closer to the Medoff-Fay concept of hours in excess of regular production work (i.e., to the 8 percent number). The 8 percent number, like the peak-to-peak interpolation work, does not account for maintenance that is shifted from high- to low-output periods.

One final point should be noted before making the comparison. The aggregate estimates are based on the assumption of constant short-run returns to labor. If there are in fact decreasing short-run returns, then, other things being equal,  $JHMIN$  will overestimate worker-hour requirements in low-output periods. This is because in off-peak output periods the values of  $\lambda$  estimated from the peak-to-peak interpolations will be lower than the true values. The Medoff-Fay results show some evidence in favor of decreasing returns to labor. The results are not very strong, however, and they do not put any stress on them. There is no obvious way

TABLE 1—ACTUAL AND PREDICTED VALUES OF  $(J \cdot H)/JHMIN$ 

Quarter	Actual	Predicted	Quarter	Actual	Predicted	Quarter	Actual	Predicted
54:I	1.022	1.020	63:III	1.026	1.026	73:I	1.000	1.009
54:II	1.021	1.026	63:IV	1.024	1.026	73:II	1.013	1.018
54:III	1.008	1.020	64:I	1.012	1.022	73:III	1.015	1.019
54:IV	1.006	1.016	64:II	1.019	1.024	73:IV	1.014	1.018
55:I	1.000	1.008	64:III	1.022	1.027	74:I	1.033	1.031
55:II	1.003	1.013	64:IV	1.026	1.029	74:II	1.031	1.031
55:III	1.011	1.015	65:I	1.018	1.019	74:III	1.042	1.038
55:IV	1.028	1.021	65:II	1.021	1.020	74:IV	1.045	1.046
56:I	1.033	1.033	65:III	1.013	1.019	75:I	1.044	1.058
56:II	1.042	1.035	65:IV	1.007	1.014	75:II	1.023	1.042
56:III	1.047	1.041	66:I	1.000	1.012	75:III	1.011	1.027
56:IV	1.043	1.037	66:II	1.008	1.021	75:IV	1.018	1.027
57:I	1.038	1.038	66:III	1.012	1.024	76:I	1.013	1.020
57:II	1.044	1.044	66:IV	1.013	1.026	76:II	1.012	1.022
57:III	1.047	1.044	67:I	1.020	1.032	76:III	1.013	1.024
57:IV	1.049	1.057	67:II	1.013	1.032	76:IV	1.011	1.023
58:I	1.054	1.071	67:III	1.015	1.030	77:I	1.000	1.013
58:II	1.047	1.062	67:IV	1.015	1.029	77:II	1.009	1.011
58:III	1.037	1.046	68:I	1.014	1.030	77:III	1.003	1.009
58:IV	1.032	1.035	68:II	1.010	1.024	77:IV	1.015	1.017
59:I	1.038	1.036	68:III	1.010	1.025	78:I	1.016	1.018
59:II	1.048	1.029	68:IV	1.015	1.029	78:II	1.017	1.006
59:III	1.055	1.034	69:I	1.028	1.028	78:III	1.019	1.011
59:IV	1.053	1.034	69:II	1.031	1.031	78:IV	1.017	1.009
60:I	1.043	1.028	69:III	1.038	1.035	79:I	1.024	1.014
60:II	1.065	1.041	69:IV	1.048	1.043	79:II	1.031	1.021
60:III	1.076	1.045	70:I	1.053	1.046	79:III	1.031	1.016
60:IV	1.084	1.052	70:II	1.051	1.045	79:IV	1.032	1.021
61:I	1.075	1.048	70:III	1.037	1.041	80:I	1.030	1.022
61:II	1.049	1.038	70:IV	1.046	1.051	80:II	1.044	1.044
61:III	1.052	1.037	71:I	1.025	1.033	80:III	1.043	1.037
61:IV	1.043	1.027	71:II	1.028	1.036	80:IV	1.045	1.031
62:I	1.044	1.028	71:III	1.024	1.037	81:I	1.030	1.019
62:II	1.043	1.028	71:IV	1.032	1.035	81:II	1.039	1.030
62:III	1.038	1.030	72:I	1.028	1.026	81:III	1.037	1.028
62:IV	1.033	1.033	72:II	1.018	1.021	81:IV	1.046	1.040
63:I	1.038	1.033	72:III	1.015	1.022	82:I	1.055	1.048
63:II	1.035	1.029	72:IV	1.011	1.017	82:II	1.050	1.041
						82:III	1.047	1.040

Note: Root mean squared error = .011. The predicted values are from a dynamic simulation that begins in 1954:I. The model consists of equations (6) and (9).  $Y$  and  $JHMIN$  ( $= Y/\lambda$ ) are exogenous.

to test the constant returns hypothesis using the aggregate data, and so it has simply been assumed to be true. One should be aware, however, that  $JHMIN$  will be biased upward if there are decreasing returns.

One thing that can be done to compare the results is simply look at the actual values of  $(J \cdot H)/JHMIN$  over the business cycle. Another is to see what the model predicts these values to be. This information is presented in Table 1. The model consists of

equations (6) and (9).  $Y$  and  $JHMIN$  ( $= Y/\lambda$ ) are exogenous. The predicted values in Table 1 are for a dynamic simulation for the 1954:I–1982:III period. The results in Table 1 show, first of all, that the model fits the data well. The predicted values are based on a dynamic simulation of 115 quarters in length, and the root mean squared error over the entire period is only .011.

Consider now the actual values in Table 1. There are two possible troughs that are rele-

TABLE 2—PREDICTED VALUES OF  $(J \cdot H)/JHMIN$  FOR ALTERNATIVE OUTPUT PATHS

Quarter	Output Change	$\frac{J \cdot H}{JHMIN}$ Change	Output Change	$\frac{J \cdot H}{JHMIN}$ Change	Output Change	$\frac{J \cdot H}{JHMIN}$ Change
78:I	-1.0	.61	-2.0	1.22	-4.0	2.48
78:II	-2.0	.97	-4.0	1.99	-8.0	4.12
78:III	-3.0	1.26	-6.0	2.58	-8.0	2.67
78:IV	-4.0	1.49	-8.0	3.10	-8.0	2.08
79:I	-4.0	1.07	-8.0	2.20		

Notes: Output Change =  $100 \cdot ((\text{new } Y / \text{old } Y) - 1)$ ;

$(J \cdot H)/JHMIN$  Change =  $100 \cdot (\text{new } J \cdot H / JHMIN / \text{old } J \cdot H / JHMIN - 1)$ .

vant for the Medoff-Fay study, the one in mid-1980 and the one in early 1982. The first survey upon which the Medoff-Fay results are based was done in August 1981, and the second (larger) survey was done in April 1982. A follow-up occurred in December 1982. The plant managers were asked to answer the questionnaire for the plant's most recent trough. For the last responses the trough might be in 1982, whereas for the earlier ones the trough is likely to be in 1980. Table 1 shows that in 1980 the percentage of excess hours reached a high of 4.5 percent in the fourth quarter. In 1982 it reached a high of 5.5 percent in the first quarter. The percentages in earlier troughs are 5.4 in 1958:I, 8.4 in 1960:IV, 5.3 in 1970:I, and 4.5 in 1974:IV.

The Medoff-Fay estimate of 8 percent is thus compared to the 4.5 and 5.5 percent values in Table 1 for the two most recent trough quarters. These two sets of results seem consistent. There are at least two reasons for expecting the Medoff-Fay estimate to be somewhat higher. First, the trough in output for a given plant is on average likely to be deeper than the trough in aggregate output, since not all troughs are likely to occur in the same quarter across plants. (In the aggregate model, other things being equal, the deeper the trough, the larger will be the predicted percentage of excess hours, and the comparison of the two sets of results has not adjusted for different size troughs.) Second, the manufacturing sector may on average face deeper troughs than do other sectors, and the aggregate estimates in Table 1 are for the total private sector, not just manufac-

turing. One would thus expect the Medoff-Fay estimate to be somewhat higher than the aggregate estimates, and 8 percent versus a number around 5 percent seems consistent with this.

With respect to the predicted values in Table 1, in 1980 the predicted percentage of excess hours reached a high of 4.4 percent in the second quarter, and in 1982 it reached a high of 4.8 percent in the first quarter. These values compare fairly closely to the actual values.

One cannot get from the Medoff-Fay results estimates of the timing of the response of excess hours to output fluctuations. This can be done, however, with the aggregate equations. The results of three experiments are reported in Table 2. These experiments were performed as follows. First, the estimated residuals were added to equations (6) and (9) and treated as exogenous. This means that when the model is solved using the actual values of  $Y$ , perfect fits are obtained for  $J$  and  $H$  (and thus  $J \cdot H$ ). Second,  $Y$  was changed and the model was solved for the new values of  $Y$ . Third, the new (predicted) values of  $J \cdot H / JHMIN$  were compared to the old (actual) values to see the response of excess hours to the output changes. The simulation period began in 1978:I. All three simulations were dynamic. For the first experiment  $Y$  was lowered (from its actual value) by 1.0 percent in the first quarter, 2.0 percent in the second, 3.0 percent in the third, and 4.0 percent in the fourth and fifth. The second experiment was the same as the first except that the decreases were twice as large. For the third experiment  $Y$  was lowered



by 4.0 percent in the first quarter and 8.0 percent in the second, third, and fourth.

The results in Table 2 show that, for the first experiment, excess hours reached a high of 1.49 percent in the fourth quarter. For the second experiment, the high was 3.10 percent in the fourth quarter. The values for the second experiment are only slightly more than twice as large as the values for the first, which means that the excess-hours response to output fluctuations is not very nonlinear with respect to the size of the changes. The response is, however, quite nonlinear with respect to the timing of the changes. For the third experiment compared to the second experiment, output was 8 percent lower by the second quarter rather than by the fourth quarter. Excess hours reached a high of 4.12 percent for the third experiment compared to a high of 3.10 percent for the second experiment.

Remember, of course, that these results are based on the particular specification of the aggregate model. If there are, for example, decreasing short-run returns to labor, then the increase in excess labor due to the fall in output will have been underestimated. Also, the model does not account for the possibility that the response of firms in eliminating excess labor is larger the larger is the fall in output. If firms begin to decrease employment drastically for very large downturns, the percentage of excess labor may actually fall as downturn size increases, and the model is not capable of capturing this. It

is unlikely that one would be able to pick up a response shift like this in the aggregate data.

### III. Conclusion

The Medoff-Fay results seem consistent with the aggregate estimates, which is further evidence in favor of the excess labor hypothesis. This hypothesis has important implications for the production function and investment literature. Much of this literature is based on the assumption that firms are always "on" their production functions. If they are not and if in fact the amount of worker hours hoarded during contractions, even after adjusting for worthwhile nonproduction work, is as much as 5 percent of total worker hours, it is not clear that estimates of production parameters and investment behavior that are based on the assumption of no hoarding are trustworthy.

### REFERENCES

- Fair, Ray C., *The Short-Run Demand for Workers and Hours*, Amsterdam: North-Holland, 1969.
- , *Specification, Estimation, and Analysis of Macroeconometric Models*, Cambridge: Harvard University Press, 1984.
- Medoff, James L. and Jon A. Fay, "Labor and Output over the Business Cycle: Some Direct Evidence," *American Economic Review*, forthcoming 1985.

# Monopoly Unionism: Note

By DANIEL LÉONARD\*

Edward Lazear (1983) presents a model in which the number of unionized workers in an industry is endogenously determined by the utility-maximizing workers themselves. Hence in his model union firms and non-union firms coexist. He assumes that the nonunion wage adjusts so that the labor market clears and full employment prevails.

The purpose of this note is to examine the possibility of unemployment in Lazear's model. Of all the possible rigidities in the labor market that could lead to unemployment, the most natural one is the existence of minimum wage legislation.<sup>1</sup> It will be shown that the imposition of a minimum wage may induce the formation of a union. Furthermore, an increase in the minimum wage will result in a higher union wage, but decrease aggregate wage income. Finally, I investigate the influence of the elasticity of demand for labor on some indicators of union power.

Henceforth it will be assumed that firms cannot pay a wage below some specified level  $\bar{W}$ . This entails some modification of Lazear's model which is now briefly presented. The variables  $W_u$  and  $W_N$  are the union wage and the nonunion wage, respectively.  $C_i$  is the cost to firm  $i$  of blocking unionization of its workforce with  $C_i \sim g(C_i)$  and  $G(C)$  the probability that firm  $i$ 's blocking cost does not exceed  $C$ ; apart from this disparity in blocking costs, all firms are identical.  $d(W)$  is the demand for labor by a firm faced with wage  $W$  and  $\Pi(W)$  is its profit outside of blocking costs. Define  $\Pi^*(W_u, W_N) \equiv \Pi(W_N) - \Pi(W_u)$  and firm  $i$  will block unionization when  $\Pi^*(W_u, W_N) > C_i$ , hence the probability that a firm will block is  $G[\Pi^*(W_u, W_N)]$ . The number of firms is  $S$

and  $R$  is the number of workers. The labor market equilibrium condition (1) of Lazear must be modified as

$$(1a) \quad S[1 - G[\Pi^*(W_u, W_N)]]d(W_u) + S[G[\Pi^*(W_u, W_N)]]d(W_N) - R \leq 0,$$

$$(1b) \quad W_N - \bar{W} \geq 0,$$

(1c) The product of the left-hand sides of (1a) and (1b) vanishes,

$$(1d) \quad W_u \geq W_N.$$

These equations define the union's opportunity locus depicted as a thick line in Figure 1. The probability that a worker will be employed at the union wage is

$$P \equiv (S/R)[1 - G[\Pi^*(W_u, W_N)]]d(W_u).$$

The probability that a worker will be employed at the nonunion wage is

$$Q \equiv \begin{cases} (S/R)[G[\Pi^*(W_u, \bar{W})]]d(\bar{W}) & \text{if } W_N = \bar{W} \\ 1 - P & \text{if } W_N > \bar{W} \end{cases}$$

The probability that a worker will be unemployed is  $(1 - P - Q)$ ; note that this is zero when  $W_N > \bar{W}$ .

Denoting the union fee by  $Z$  each worker chooses  $W_u$  and  $W_N$  to maximize

$$(2) \quad K = P(W_u - Z) + Q(W_N) + (1 - P - Q)(0),$$

subject to (1).

In the case where (1b) is not binding while (1a) is, the analysis of Lazear applies without change. Hereafter I analyze the case where (1a) is not binding while (1b) is; then the worker's problem reduces to choosing  $W_u$  to

\*Department of Econometrics, University of New South Wales, P.O. Box 1, Kensington, N.S.W. 2033, Australia. I am grateful to Edward Lazear for a constructive suggestion.

<sup>1</sup>Jacob Mincer (1981) addresses the minimum wage problem in a somewhat similar vein.

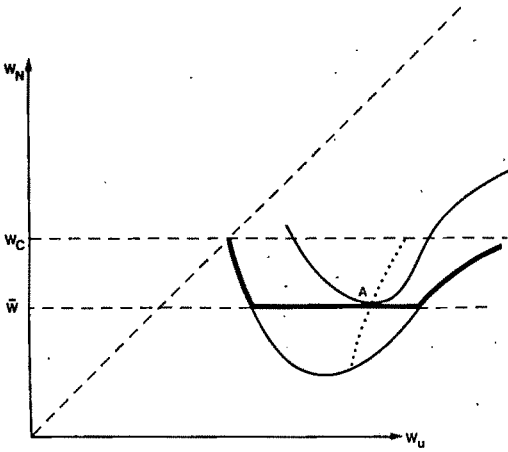


FIGURE 1

maximize

$$\begin{aligned}
 K &= (S/R)[1 - G[\Pi^*(W_u, \bar{W})]]d(W_u) \\
 &\quad \times (W_u - Z) + (S/R) \\
 &\quad \times [G[\Pi^*(W_u, \bar{W})]]d(\bar{W})\bar{W}, \\
 K &= (S/R)[G[\Pi^*(W_u, \bar{W})]] \\
 &\quad \times [d(\bar{W})\bar{W} - d(W_u)(W_u - Z)] \\
 &\quad + d(W_u)(W_u - Z)].
 \end{aligned}$$

A solution to this problem must satisfy the first-order condition,

$$\begin{aligned}
 (dG/dW_u)[d(\bar{W})\bar{W} - d(W_u)(W_u - Z)] \\
 + (1 - G)[d'(W_u)(W_u - Z) + d(W_u)] = 0.
 \end{aligned}$$

But  $dG/dW_u = G'(\partial\Pi^*/\partial W_u) = G'd(W_u)$  from duality theory, and the first-order condition can be stated as

$$\begin{aligned}
 (3) \quad (1 - G)[d'(W_u)(W_u - Z) + d(W_u)] \\
 - G'd(W_u)[d(W_u)(W_u - Z) - d(\bar{W})\bar{W}] = 0
 \end{aligned}$$

This maximum is situated at point *A* in Figure 1. Condition (3) can be interpreted as follows: the first term is the product of the

rate of increase in union revenue per firm, and of the proportion of unionized firms; the second term is the product of the additional total wage payment made by a union firm over that made by a nonunion firm, and of the rate of increase in the proportion of nonunion firms. At the optimum the expected gain from increasing  $W_u$  (while remaining employed by a union firm) is exactly offset by the expected loss due to increased militancy of firms and the subsequent increase in the number of nonunion firms.

Let us now consider the effect of an increase in the minimum wage rate on the level of the union wage. At a maximum the sign of  $dW_u/d\bar{W}$  is given by the partial derivative of (3) with respect to  $\bar{W}$ :

$$\begin{aligned}
 A &\equiv G'd(\bar{W})[d'(W_u)(W_u - Z) + d(W_u)] \\
 &\quad + G''d(\bar{W})d(W_u)[d(W_u)(W_u - Z) \\
 &\quad - d(\bar{W})\bar{W}] \\
 &\quad + G'd(W_u)[d'(\bar{W})\bar{W} + d(\bar{W})].
 \end{aligned}$$

Using (3) we obtain,

$$\begin{aligned}
 A &= [d'(W_u)(W_u - Z) + d(W_u)]d(\bar{W}) \\
 &\quad \times (1 - G)[G'/(1 - G) + G''/G'] \\
 &\quad + G'd(W_u)[d'(\bar{W})\bar{W} + d(\bar{W})].
 \end{aligned}$$

The first and last brackets in *A* are marginal revenues to be extracted from a firm by workers; clearly wage claims (and a fortiori  $\bar{W}$ ) will not be pushed past the point where they yield negative returns. Therefore a sufficient (but not necessary) condition for *A* to be positive is  $[G'/(1 - G) + G''/G'] \geq 0$ . This condition restricts only the distribution of the costs of union blocking among firms; if it is satisfied we say that the distribution of these costs is "regular."<sup>2</sup>

<sup>2</sup>It is clearly possible to find distributions that violate this regularity condition so that the result may not hold. However, this is only possible when *g* is decreasing and thus, for unimodal distributions, restricts only the upper tail. The mildness of the condition can be better grasped

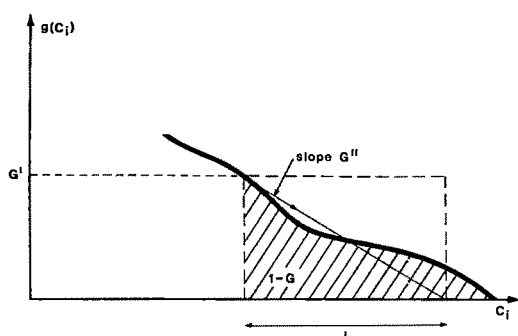


FIGURE 2

**PROPOSITION 1:** *If blocking costs are regularly distributed, the imposition of or an increase in the minimum wage will raise the union wage, decrease employment in both sectors, and decrease aggregate wage income (hence the expected income of each worker). The larger the increase in the minimum wage, the stronger are these effects.*

Most of Proposition 1 follows directly from the above argument. The reason why aggregate wage income decreases with  $\bar{W}$  is that  $K$  is the average wage income; therefore any further tightening of the minimum wage constraint lowers the attainable maximum  $K$ .

Since this result holds in all cases where the minimum wage constraint is binding, we can draw the locus of equilibrium points as  $\bar{W}$  rises; it is depicted by the dotted line in Figure 1. Proposition 1 implies that minimum wage legislation has little to recommend it: firms' profits, average worker's income, and output all fall as a result. The only unambiguous beneficiaries are union workers still employed after the change (non-union workers who are still employed may

have been union workers previously). The argument often proffered in defense of a minimum wage that it is the lowest "living" wage is not tenable in this model in view of the possibility that some workers are unemployed.

Lazear correctly points out (p. 634) that if the critical indifference curve (which yields the same utility as that derived in competitive equilibrium) does not intersect the opportunity locus, a union equilibrium fails to exist, that is, all workers receive the competitive wage (denoted by  $W_c$  in Figure 1). The introduction of a minimum wage has a powerful implication for this existence question. Clearly, if  $\bar{W}$  is large enough so that the critical indifference curve intersects the line  $W_N = \bar{W}$ , a union equilibrium will occur.

**PROPOSITION 2:** *The imposition of a high enough minimum wage will result in the formation of a union sector in an industry with a previously competitive labor market.*

Obviously, the policy implications of these results depends on the empirical validity of the model; thus it is tempting to attempt to test it by examining whether the unionized sector has grown after the introduction of minimum wage legislation in various states. It is comforting to note that unions have typically been staunch supporters of a minimum wage, as predicted by this note.

One finding reported by Lazear is that "Contrary to Marshall, inelasticity of the demand for labor does not imply an increase in union power as measured by either an increase in union membership or the union/nonunion wage differential" (p. 631). I now show that when the minimum wage constraint is binding, the wage differential unmistakably increases with the inelasticity of the demand for labor. The explanation for this reversal lies in the fact that while in Lazear's model the secondary effect of an increase in the union wage was to reduce the nonunion wage thereby making the net effect ambiguous (see p. 636), no such occurrence is possible when the nonunion workers receive the minimum wage.

Let us consider small changes in the elasticity of demand for labor, denoted by  $E > 0$ ,

with the help of Figure 2. The length  $L$  is defined by  $L = -G'/G''$  and

$$G'/(1-G) + G''/G' = (G'L/(1-G) - 1)/L \geq 0.$$

Thus the regularity condition requires that the area under the upper tail does not at any point exceed the area of the rectangle that has the tangent to  $g$  as its diagonal. It is obvious that this is true of concave, linear, or mildly convex functions.

starting from the equilibrium position defined by (3) which can be rewritten as

$$(4) \quad (1-G)[1-E(W_u-Z)/W_u] - G'[d(W_u)(W_u-Z)-d(\bar{W})\bar{W}] = 0.$$

At a maximum the sign of  $dW_u/dE$  is given by the partial derivative of (4) with respect to  $E$ :

$$B \equiv -(1-G)(W_u-Z)/W_u < 0.$$

Therefore  $dW_u/dE < 0$ .

However, the effect of a higher wage on other indicators of union power is negative as shown:

$$\begin{aligned} dP/dE &= (dP/dW_u)(dW_u/dE) \\ &= (S/R)[d'(W_u)(1-G) \\ &\quad - G'(d(W_u))^2](dW_u/dE) > 0, \end{aligned}$$

and

$$\begin{aligned} dQ/dE &= (dQ/dW_u)(dW_u/dE) \\ &= (S/R)G'd(W_u)d(\bar{W})(dW_u/dE) \\ &< 0. \end{aligned}$$

Let total net pay of union workers be denoted by  $T \equiv RP(W_u - Z)$ .

$$T = S(1-G)d(W_u)(W_u-Z)$$

$$\begin{aligned} dT/dE &= (dW_u/dE)S[(1-G) \\ &\quad \times [d'(W_u)(W_u-Z) + d(W_u)] \\ &\quad - G'(d(W_u))^2(W_u-Z)]. \end{aligned}$$

Using (3) this simplifies to

$$\begin{aligned} dT/dE &= -SG'd(W_u)\bar{W}d(\bar{W})(dW_u/dE) \\ &> 0. \end{aligned}$$

**PROPOSITION 3:** *When nonunion workers receive the minimum wage, the inelasticity of the demand for labor increases the union wage, decreases the proportion of union workers, and increases that of nonunion workers. It also decreases the total net pay of union workers.*

Hence we would expect to find fewer union firms and a higher union/nonunion wage differential in industries with a more inelastic labor demand.

Although these predictions are quite clear, they affect union power in opposite ways. On the whole, I would argue that a smaller total net pay for the union membership means a weaker union and beg to differ from Marshall in this case.

## REFERENCES

- Lazear, Edward P., "A Competitive Theory of Monopoly Unionism," *American Economic Review*, September 1983, 83, 631-43.  
Mincer, Jacob, "The Economics of Wage Flows," Working Paper No. 804, National Bureau of Economic Research, November 1981.

# Why Everything Takes 2.71828... Times as Long as Expected

By PHILIP MUSGROVE\*

It is widely observed, and almost as widely lamented, that everything takes longer than one expects. However, most attempts to explain why deadlines are missed and budgets overrun go no farther than Murphy's (n.d.) often-quoted aphorism. Blaming the phenomenon on unrealistic expectations, as Handtvefer (1982) does, cannot explain why expectations are not revised after repeated disappointment. The problem presents both a theoretical challenge to economic science and an issue of great practical importance; a procedure for predicting delays could save a lot of money and frustration. In the absence of constraints on the time available for a job, it turns out that the ratio of time taken to time expected tends to  $e = 2.71828...$  for a job consisting of an infinite number of steps. Shorter jobs exceed the expected time by ratios less than  $e$  but never less than 2. Constraints on time, when the time available is less than what is expected to be needed for completion, only make matters worse.

## I. Delays in Steps and in Jobs

A "job" is just something one wants to get done—and can tell whether it has been finished or not. A "step" is a physically essential part of a job, performed in sequence with other steps; completing the last step means finishing the job. To avoid considering intervals between steps, a step is not regarded as finished until the next step is begun.

The first question then is, how does delay in a step affect the job of which the step is a part? Procrastinateur (1971) appears to have been the first to show that a delay in one step introduces an equal proportional, rather than absolute, delay in the entire project. Study-

ing 283 large civil engineering projects,<sup>1</sup> in each of which one step triggered delay in the job, he found that a one-month step delay could slow down the project by as much as two years, even when nothing else went exogenously wrong. I have obtained very similar results using a large sample from the *PAO Register* (1969 et seq.), which includes military as well as civilian projects discriminated by presidential administration, cabinet agency or contractor responsible, and the state in which the project occurred. None of these variables is significant.<sup>2</sup>

These results can be summarized in

**PROPOSITION 1:** *If  $S$  and  $P$  are step and project time, respectively, and  $DS$  and  $DP$  are step and project delays, then*

$$1 + DP/P = \prod_{i=1}^N (1 + DS_i/S_i)$$

where  $i = 1, 2, \dots, N$  are the project steps.

## II. Number of Steps and Step Delays

The finding that job delays are proportional to step delays is initially counterintuitive, but is plausible once one considers that the delayed step interrupts the schedule of work, causes overtime or stretch-outs to avoid overtime, and even produces delays in previously executed steps which have to be tested or repeated to make sure they do not suffer the same error which caused the delay. Procrastinateur's research also suggests that job delay does not depend on where in the project the delayed step or "foul-up" occurs,

<sup>1</sup>All of them, to be sure, designed and executed by Frenchmen.

<sup>2</sup>When highway projects alone are studied, the dummy for Massachusetts is almost significant ( $t = 1.83$ ).

\*The Brookings Institution, 1775 Massachusetts Avenue, NW, Washington, D.C. 20036.

contradicting Harnischfeger's (1976) hypothesis that delays hurt most if they happen near the beginning or end of a project.<sup>3</sup> A still more puzzling result is that the complexity or differentiation of the steps does not seem to matter: this contradicts the intuitive notion that a step is a quite arbitrary element of a project, and that complex steps can be made into simple ones by subdivision. This should permit better control and less delay, but if simple and complicated steps are equally dangerous as sources of delay, nothing can be gained by adding to the number of steps. A project such as raising a pyramid, in which each step consists of placing a single stone but some are harder to seat than others, might offer a test of this issue. Unfortunately, Bloch's (1948) estimates of delays due to fractures and accidents at Giza are hotly disputed by other scholars, and anyway are too vague for quantitative analysis.<sup>4</sup>

Procrastinateur did not include the number of steps in his analysis, but when it is included, the apparent paradox disappears. Smythe (1976), studying the limits on the division of labor in manufacturing, discovered that step delay tends to be inversely proportional to the number of steps in a process, confirming that subdivision of labor yields a gain at the level of individual steps. Her results give

**PROPOSITION 2:**  $E(DS/S) = 1/N$ , with  $\text{var}(DS/S)$  of order  $1/N^2$ .

Combining Propositions 1 and 2 then establishes

**THEOREM 1:** *As a job is continuously subdivided into steps, the time actually taken to complete it tends to  $e$  times the time anticipated for its completion.*

<sup>3</sup>Harnischfeger's hypothesis, while incorrect for delays, may still be valid for "step problems" of other types—for example, if the job is an airplane flight, with takeoff the first step and landing the last one.

<sup>4</sup>Bloch sometimes relies on the number of days the Pharaoh spent sulking, which, even if correctly reported by the scribes, is a poor proxy for delay in construction.

#### PROOF:

It is straightforward: substitution of Proposition 2 into Proposition 1 gives:

$$\begin{aligned} \text{Time Required/Time Expected} \\ &= R_N = (1 + 1/N)^N \\ \text{and } \lim R_N &= e, \text{ as } N \rightarrow \infty. \end{aligned}$$

Passing to the limit removes the arbitrariness in the definition of a step. The ratio nevertheless converges: a job of arbitrarily many steps does not take forever to complete, in accord with experience.<sup>5</sup> It is an immediate corollary that even a job of only one step takes on average twice as long to finish as expected, since  $N=1$  means  $R_N = 2$ .

#### III. Time Required and Time Available

Suppose  $A$  is the total time available for a job (exogenously determined by the "boss"), and that the average anticipated time required for a step is  $T$ , so that  $NT$  is the total time the job is expected to take. Define  $m = NT/A$  as the share of the available time that the job is expected to use. If  $N$  and  $A$  tend to infinity at the same rate, holding  $m$  constant, then the probability of actually completing  $x$  steps in an interval  $T$  is described by the Poisson distribution,

$$p(x) = e^{-m} m^x / x!$$

where  $x$  is a random variable because the steps vary in length or difficulty. The rapid decline in  $p(x)$  as  $x$  increases reflects the unlikelihood of "catching up" a delayed step by completing an additional step in a later interval.<sup>6</sup> The terms  $m^x/x!$  sum to  $e^m$  over all nonnegative values of  $x$ .

If  $m=1$ , so that the time appears to be just adequate, the project will in fact take 2.71828... times as long to complete as anticipated. What happens if the time available *ex ante* appears to be too short? If  $m=2$ , for

<sup>5</sup>So-called "interminable" jobs are the proper domain of Murphy's Law.

<sup>6</sup>A complete theory of personal stress reduction is built on this by Orff (1969).

example, the ratio of time taken to time expected becomes  $e^2$  or 7.3891...: trying to squeeze a project into half the required time makes it take not twice as long to finish as was anticipated, but nearly eight times as long. This explains Shaughnessy's (1937) finding that attempts to speed up projects by unrealistic deadlines actually end by slowing them down.<sup>7</sup> Letting  $C$  be the time taken to complete a project, the foregoing establishes

**THEOREM 2:**  $R = C/NT \rightarrow \exp(NT/A)$ , as  $N, A \rightarrow \infty$ ; for  $T$  constant.

It is a corollary of this theorem that a job can actually be completed in exactly the time expected—a result which Theorem 1 does not permit—but only on condition that it be expected to use a negligible fraction of the available time. The proof is direct:  $R=1$  requires  $\exp(NT/A)=1$ , from which  $NT/A=0$ . This result also accords with experience: the only jobs finished on time are those for which nobody is in a hurry.

<sup>7</sup>Shaughnessy's results for Soviet labor productivity are violently disputed by Vodkapiu (1937), who—curiously—published first. Subsequent evidence, however, tends to vindicate Shaughnessy.

## REFERENCES

- Bloch, Pyramus, "Y' Raise Sixteen Stones," *Review of Archaeoeconomics*, Special Issue on Ancient Public Works, May 1948, 22, 68–94.
- Handtvefer, Luke, "Micro-Rational Expectations, or Learning by Doing (Wrong)," *Ausgezeichneter Stiftung*, June 1982, 3, 1–11.
- Harnischfeger, Harold, "Project Delays, Cost Overruns and the Right Way to Run a Railroad," *Journal of Marginal Management*, Spring 1976, 41, 37–50.
- Murphy, N. M., *If Anything Can Go Wrong*, Secaucus: Nullo Modo Press, n.d.
- Orff, Hans, "One Thing at a Time," *Journal of Statistics and Self-Fulfillment*, February 1969, 1, 3–28.
- Procrastinateur, Jean-Jacques, "Les Retards dans Les Grands Projets de Construction Civile," *Annales de l'Académie des Imperfections*, Juillet 1971, 83, 115–32.
- Shaughnessy, Brendan O., "Stakhanovism, Speed-ups and Reduced Productivity," *Soviet Management Studies*, April 1937, 10, 68–101.
- Smythe, Adele, "The Precision of Labor is Limited by the Extent of the Market," *American Economic Journal*, Smith Bicentennial Number, September 1976, 138, 25–62.
- Vodkapiu, Ivan Akakievitch, "Shto Znayet Shaughnessy?," *Akademia Naukonomika CCCP*, March 1937, 19, 1–3.
- U.S. Partial Accounting Office, *Register of Federally Financed Projects*, Vol. XLIII, *Contract and Actual Dates and Costs of Completion*, Washington: USGPO, 1969 et seq.
- Bloch, Pyramus, "Y' Raise Sixteen Stones," *Review of Archaeoeconomics*, Special Issue



# The Design of an Optimal Insurance Policy: Note

By GEORGE BLAZENKO\*

In an article in this *Review*, Arthur Raviv (1979) examines Pareto optimal insurance policies when an insurer incurs settlement costs  $C$  induced by indemnity " $I$ " for loss  $x$ . Raviv's main result is that a necessary and sufficient condition for the Pareto optimal deductible to equal zero is  $C'(I) = 0$ . This implies that deductible policies give the best tradeoff between risk sharing and economizing on costly claim settlements. Since in practice these costs are significant, the theorem is of considerable importance. Among others, this has been recognized by Robert Townsend (1979), Michael Brennan and Ray Solanki (1981), David Mayers and Clifford Smith (1981), Gur Huberman, Mayers, and Smith (1983), Harris Schlesinger (1981), and Stuart Turnbull (1983). The theorem is correct, but Raviv's proof is not.

In this note a corrected proof for the theorem is given. The corrected proof is important in itself because it allows for a generalization to a greater variety of transactions costs than has previously been considered (see my 1984 paper for details). Section I develops the setting for the problem and the notation to be subsequently used. Raviv's error and the corrected proof are presented in Section II.

## I. Insurance with Costly Claim Settlement

The insured faces a random loss  $x$ ,  $0 \leq x \leq T$ , with density  $f(x) > 0$ . Insurance indemnity is given by the schedule

$$(1) \quad 0 \leq I(x).$$

Costs of claim settlement are given by  $C[I(x)]$  with  $C' \geq 0$ . This cost represents a deadweight loss relative to both insurer and

insured. For simplicity, it is assumed that  $C[0] = 0$ .

The insurer's final wealth is  $W_0 + P - I(x) - C[I(x)]$ , where  $W_0$  is initial wealth and  $P$  is the premium. The insured's final wealth is  $\omega - P - x + I(x)$ , where  $\omega$  is the insured's initial wealth.

Twice differentiable utility functions for insurer and insured are  $V$  and  $U$ , respectively, both concave increasing. To find Pareto optimal contracts the insured's expected utility is maximized subject to the insurer reaching a required utility level.

$$(2) \quad \max_{P, I(\cdot)} \int_0^T U\{\omega - P - x + I(x)\} f(x) dx,$$

subject to (1) and

$$(3) \quad \int_0^T V\{W_0 + P - I(x) - C[I(x)]\} \times f(x) dx \geq K \geq V(W_0).$$

The Hamiltonian is

$$(4) \quad U\{\omega - P - x + I(x)\} f(x) + \lambda V\{W_0 + P - I(x) - C[I(x)]\} f(x),$$

where  $\lambda$  is a constant. Necessary conditions for a maximum with respect to indemnity are,

$$(5) \quad U'\{\omega - P - x\} - \lambda V'\{W_0 + P\} \times (1 + C'[0]) \leq 0 \quad I^* = 0,$$

$$(6) \quad U'\{\omega - P - x + I^*\} - \lambda V'\{W_0 + P - I^* - C[I^*]\} (1 + C'[I^*]) = 0 \quad I^* > 0.$$

Assuming the Hamiltonian is concave in  $I$ , necessary conditions are sufficient for a maximum. The relation,  $P(\bar{x})$ , between the de-

\*Accounting Group, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada.

ductible  $\bar{x}$  and the premium  $P$  is

$$(7) \quad U'\{\omega - P - \bar{x}\} - \lambda V'\{W_0 + P\}(1 + C'[0]) = 0.$$

Equation (6) implicitly defines  $I^*$  as a function of  $x$  and  $P$ , but since  $P$  is an implicit function of  $\bar{x}$  through (7),  $I^*$  can be considered a function of  $x$  and  $\bar{x}$ .

## II. Deductible Policies

Raviv's aim is to find  $\partial P / \partial \bar{x}$  given that the insured's expected utility is constant and  $\partial P / \partial \bar{x}$  given that the insurer's expected utility is constant, and show that the former is greater than the latter. If this is true, the reduction in premium the insured requires to maintain utility for an increase in deductible is less than the maximum premium reduction the insurer is willing to accept. In such a case it is possible to increase utilities of both by increasing the deductible. The implicit function theorem is used to find the derivatives. These depend upon the optimal indemnity schedule  $I^*$ , which in turn depends upon the loss and the deductible. However,  $I^*(x, \bar{x})$  and  $\partial I^*(x, \bar{x}) / \partial \bar{x} = \partial I^* / \partial P \cdot dP(\bar{x}) / d\bar{x}$ , which appear in the desired derivatives encompass equation (7) which determines the relation  $P(\bar{x})$ . This means that when writing insurer and insured expected utilities as functions of the indemnity schedule  $I^*(x, \bar{x})$  the relation  $P(\bar{x})$  between the premium and the deductible is assumed. To seek a further relationship between the premium and deductible holding expected utilities constant is inappropriate. Insurer and insured expected utilities based on the indemnity schedule  $I^*(x, \bar{x})$  must be treated as functions of  $\bar{x}$  alone, the premium  $P$  and the deductible  $\bar{x}$  are not free to vary independently.

### A. Corrected Proof

From (7)

$$(8) \quad dP/d\bar{x} = -R_u(\omega - P - \bar{x}) / [R_u(\omega - P - \bar{x}) + R_v(W_0 + P)] \leq 0,$$

where  $R_u$  and  $R_v$  are the Arrow-Pratt indices of risk aversion. The insured's expected utility with a deductible policy is

$$U^*(\bar{x}) = \int_0^{\bar{x}} U\{\omega - P(\bar{x}) - x\} f(x) dx + \int_{\bar{x}}^T U\{\omega - P(\bar{x}) - x + I^*(x, \bar{x})\} f(x) dx.$$

The derivative with respect to  $\bar{x}$  is

$$-P'(\bar{x}) \int_0^{\bar{x}} U'\{\omega - P(\bar{x}) - x\} f(x) dx + \int_{\bar{x}}^T U'\{\omega - P(\bar{x}) - x + I^*(x, \bar{x})\} \times [-P'(\bar{x}) + \partial I^* / \partial \bar{x}] f(x) dx.$$

The insurer's expected utility is

$$V^*(\bar{x}) = \int_0^{\bar{x}} V\{W_0 + P(\bar{x})\} f(x) dx + \int_{\bar{x}}^T V\{W_0 + P(\bar{x}) - I^*(x, \bar{x}) - C[I^*]\} f(x) dx.$$

The derivative with respect to the deductible is

$$V'\{W_0 + P(\bar{x})\} P'(\bar{x}) \int_0^{\bar{x}} f(x) dx + \int_{\bar{x}}^T V'\{W_0 + P(\bar{x}) - I^*(x, \bar{x}) - C[I^*]\} \times [P'(\bar{x}) - (1 + C') \partial I^* / \partial \bar{x}] f(x) dx.$$

Problem (2) is completed by maximizing

$$(9) \quad U^*(\bar{x}) + \lambda V^*(\bar{x})$$

with respect to  $\bar{x}$ ; this gives the Pareto optimal deductible. Constraint (3) serves to determine  $\lambda$ .

The derivative of (9) evaluated at  $\bar{x} = 0$  is

$$\begin{aligned} & -P' \int_0^T [U'\{\omega - P - x + I^*\} \\ & \quad - \lambda V'\{W_0 + P - I^* - C\}] f(x) dx \\ & + \int_0^T [U'\{\omega - P - x + I^*\} \\ & \quad - \lambda V'\{W_0 + P - I^* - C\} (1 + C')] \\ & \quad \times \partial I^* / \partial \bar{x} \cdot f(x) dx. \end{aligned}$$

From (6) the second term is zero. Also from (6), and assuming  $C' > 0$ , the first term in square brackets is positive. This means the entire first term is positive and therefore the Pareto optimal deductible is not zero. If  $C' = 0$ , both terms are zero and the Pareto optimal deductible is zero.

If (9) is maximized at  $\bar{x} = T$ , a risk-sharing arrangement that makes both parties better off is not possible. In this case the costs of claim settlement overwhelm the advantages of risk sharing.

#### REFERENCES

- Blazenko, George, "Optimal Insurance Policies," Working Paper, Faculty of Commerce and Business Administration, University of British Columbia, 1984.
- Brennan, Michael J. and Solanki, Ray, "Optimal Portfolio Insurance," *Journal of Financial and Quantitative Analysis*, September 1981, 16, 279-90.
- Huberman, Gur, Mayers, David and Smith, Clifford W., "Optimal Insurance Policy Indemnity Schedules," *Bell Journal of Economics*, Fall 1983, 14, 415-26.
- Mayers, David, and Smith, Clifford W., "Contractual Provisions, Organizational Structure, and Conflict Control in Insurance Markets," *Journal of Business*, July 1981, 14, 407-34.
- Raviv, Arthur, "The Design of an Optimal Insurance Policy," *American Economic Review*, March 1979, 69, 84-96.
- Schlesinger, Harris, "The Optimal Level of Deductible in Insurance Contracts," *Journal of Risk and Insurance*, September 1981, 48, 465-81.
- Townsend, Robert M., "Optimal Contracts and Competitive Markets with Costly State Verification," *Journal of Economic Theory*, October 1979, 21, 265-93.
- Turnbull, Stuart M., "Additional Aspects of Rational Insurance Purchasing," *Journal of Business*, April 1983, 56, 217-29.

# A Further Comment on Preemptive Patenting and the Persistence of Monopoly

By JONATHAN A. K. CAVE\*

Several recent papers in this *Review* have raised the question of whether the patent system can lead to inefficiencies in the innovative process by encouraging inefficient incumbent firms to preempt more efficient rivals. (See Richard Gilbert and David Newbery, 1982; 1984, and Stephen Salant, 1984.) In the course of this discussion, the focus has shifted away from the interesting question of whether innovation will be done efficiently to the tangential issue of whether an incumbent firm or a rival will do the innovation. I shall argue that any such inefficiency of innovation is limited by the possibility of bargaining, and that the issues of which markets are allowed to function and which party does the innovation, raised in the Reply by Gilbert and Newbery (1984) to the Comment by Salant, are somewhat beside the point.

Salant argued that the possibility of post-innovation bargaining between the incumbent and a more efficient entrant limits the profitability of inefficient preemption. Gilbert and Newbery point out that the incumbent may be able to purchase the *R&D* technology of the entrant at a price, which means that the incumbent will still develop the innovation. I wish to make two simple points. The first is that the transactions cost of purchasing an alternative *R&D* technology serves as an upper bound on the cost disadvantage of the incumbent firm, so its introduction reinforces Salant's contention that the adverse efficiency implications of the patent system are less than they appeared in Gilbert and Newbery's 1982 article. The second point is that the possible purchase of the efficient *R&D* technology by the incumbent limits the extent to which innovation by the incumbent merits the epithet "inefficient."

As I understand it, Salant's contention runs something like the following. An incumbent facing the possibility of a patent race has the following choice: he can wait for an entrant to win the race, and then strike a bargain with that entrant; or he can spend enough money to preempt the entrant, this being the amount the entrant spends plus the "disadvantage"  $\Delta$ .

Suppose that a successful entrant would earn an amount  $X$  from the bargain; competition among entrants will ensure that each serious contender is "bidding"  $X$ , so that the incumbent firm must spend at least  $X + \Delta$  to preempt. Assuming that monopoly profits are  $M$ , the value of preemption to the incumbent is

$$(1) \quad M - X - \Delta.$$

On the other hand, the bargain may not be costless. Let us assume that it costs  $T$  to consummate the bargain. Assuming that the result of the bargain is monopoly profits to the industry, the monopolist will get

$$(2) \quad M - X - T.$$

Salant's point is that *R&D* will be done by the inefficient incumbent iff the quantity in (1) exceeds the quantity in (2), or, equivalently, iff

$$(3) \quad T \geq \Delta,$$

which seems unsurprising. In other words, the opening of a costly market between the incumbent and the entrant places a bound on the inefficiency characterizing the innovative process. A firm with a large cost disadvantage would not choose to preempt; instead it would choose the more profitable path of waiting to see which entrant wins the race, and then striking a bargain with that entrant.

\*The Rand Corporation, 1700 Main Street, Santa Monica, CA 90406. I thank Robert Clower and an anonymous referee for editorial assistance.

In their Reply, Gilbert and Newbery differ heatedly with this conclusion. They point out that the market opened by Salant is just one of several markets whose closure or inefficient (costly) functioning is responsible for their result. They mention the bargain over the fruits of the patent race that Salant refers to, with the difference that the transactions cost  $T$  is renamed  $\Gamma_2$ . In addition, they mention another possibility: the incumbent could obviate his cost disadvantage by purchasing the superior  $R\&D$  technology that, in their model, is the sole reason for the entrants' cost advantage over the incumbent. This transaction has an associated transactions cost that Gilbert and Newbery denote  $\Gamma_1$ .

Before going further, I should mention that the source of this superior technology and the nature of the transactions cost  $\Gamma_1$  are important. If the incumbent does acquire the new technology and thereby wins the race, the innovation has at least been developed with the efficient technology. Therefore, preemption is not necessarily inefficient. On the other hand, unless the new technology is purchased from one of the most efficient entrants and  $\Gamma_1$  is a pure rent to that entrant, the cost to the industry of realizing the innovation is higher than it might otherwise have been.

Gilbert and Newbery allow for additional unmeasured imperfections in carrying out either of the bargains referred to, so that total industry profits from the post-innovation bargain between the incumbent and the successful entrant may be less than  $M - \Gamma_2$ . In addition, the incumbent who succeeds in beating an entrant spending  $X$  may spend more than  $X + \Gamma_1$ . However, the authors offer no explanation for these leakages, and I shall neglect them.

In any event, if the incumbent buys the new technology (for  $X + \Gamma_1$ ) and beats the entrant, he gets

$$(4) \quad M - X - \Gamma_1.$$

If the incumbent preempts the entrant using his old (inefficient)  $R\&D$  technology, he gets

$$(5) \quad M - X - \Delta.$$

On the other hand, if he loses and bargains with the entrant, he gets

$$(6) \quad M - X - \Gamma_2.$$

In this presentation, preemption will be worthwhile iff the transactions cost associated with bargaining over the patent exceeds either the transactions cost associated with purchasing the new  $R\&D$  technology or the incumbent's cost disadvantage:

$$(7) \quad \Gamma_2 \geq \min\{\Gamma_1, \Delta\}.$$

$\Gamma_2$  and  $T$  are manifestly identical; the substantive issue concerns the relation between  $\Gamma_1$  and  $\Delta$ .<sup>1</sup>

If (7) does not hold, the innovation will be developed (efficiently) by an entrant using the better  $R\&D$  technology. If (7) does hold, and if  $\Delta > \Gamma_1$ , the innovation will be developed by the incumbent, but the incumbent will be using the new technology. If the issue is the efficiency of innovation and not the question of who does the innovation, our attention shifts to the nature of the transactions cost  $\Gamma_1$ . To the extent that it represents a redistribution of profits within the industry, it does not affect the conclusion that innovation is still efficient. In any event, it places an upper bound on the inefficiency of innovation. Finally, if (7) holds, and  $\Delta \leq \Gamma_1$ , innovation will preempted by the inefficient incumbent.

Even if we take the position that the entire  $R\&D$  transactions cost  $\Gamma_1$  represents an inefficiency, we are left with the conclusion that the amount lost is limited. The introduction of the additional transaction has only served to weaken the argument against the patent system. On the other hand, the separation of this transactions cost from the incumbent's "cost disadvantage" requires some justification. As economists, we should

<sup>1</sup>In their Reply, Gilbert and Newbery focus on the "relative transactions costs" as a determinant of preemption. While I have argued that it is the cost disadvantage of the incumbent that must be compared to the transactions cost of bargaining over the patent, it seems quite implausible that the latter should exceed the cost of bargaining over the "pig in a poke" of a new  $R\&D$  technology.

be sufficiently comfortable with the concept of opportunity cost to recognize that in the absence of differential uncertainty about the outcome of innovative activity  $\Gamma_1$  is itself an upper bound on the cost disadvantage of the incumbent firm. After all, there are many ways of doing *R&D*, one of which is to purchase and use a new technology. A rational firm would certainly use the least expensive of these. This implies that  $\Gamma_1 \geq \Delta$ , and reduces equation (7) to Salant's equation (3). Thus, the transactions cost introduced by Gilbert and Newbery is irrelevant to the efficiency and preemption issues.

More importantly, while either  $\Gamma_1$  or  $\Delta$  may be relatively large, the meaning of (7) is that they will only result in efficiency losses if one of them is less than the transactions cost of patent transfer,  $\Gamma_2$ , which in reality is often fairly small.

All of the papers address themselves to the question of whether the possibility of preemption will lead to inefficient *R&D* expenditures. While it is difficult to discuss this sensibly in the context of a model that abstracts from demand and from the real resources consumed by unsuccessful competitors in a patent race, some lessons can be drawn. One of these, which I understood as being the point of Salant's Comment, is that the possibility of bargaining places an upper

bound on the inefficiency created by preemptive patenting. Gilbert and Newbery's Reply seems to miss this point entirely. The economically relevant issue is not whether preemption can occur, but whether the possibility of preemption has negative efficiency implications. In their attempt to deny Salant's contention that the possibilities for inefficient preemption are not as great as might appear in the absence of bargaining, Gilbert and Newbery inadvertently reinforce the implication that the inefficiency associated with innovation is bounded by the possibility of bargaining, regardless of who wins the patent race.

## REFERENCES

- Gilbert, Richard J., and Newbery, David M. G., "Preemptive Patenting and the Persistence of Monopoly," *American Economic Review*, June 1982, 72, 514-26.
- \_\_\_\_\_ and \_\_\_\_\_, "Preemptive Patenting and the Persistence of Monopoly: Reply," *American Economic Review*, March 1984, 74, 251-53.
- Salant, Stephen W., "Preemptive Patenting and the Persistence of Monopoly: Comment," *American Economic Review*, March 1984, 74, 247-50.

## A Curse on Several Houses

By BORIS P. PESEK\*

Suppose that the consumer has the choice among (a) a zero-yield and zero-risk asset, (b) a high-yield and excessively risky asset, (c) a low-yield and zero-risk asset, and (d) some real asset. If we restrict the consumer to the first two, he will select the first one, called "speculative balances." In our world in which the consumer has all four choices, he will never select these balances. One economist (see my 1976 paper) wasted almost two pages of the *Journal of Economic Literature* proving that a consumer will prefer a positive risk-free income to zero income. Three economists (Winston Chang, Daniel Hamberg, Junichi Hirata, 1983) wasted eight pages and some matrix algebra on a renewed effort to persuade the profession of the same thing. Linear extrapolation suggests that a renewed effort will become due in 1990.

But the problem is more serious. Given that the demand for speculative balances must be zero, Chang et al. ask (p. 46) why so many texts analyze the demand for them. I would add that these texts go on to raise the specter of a price-theoretic monstrosity—an infinite demand for speculative balances ("the liquidity trap") and its devastating consequences for monetary policy. This is not unusual. Texts give much space to the theory that current consumption is determined by unmeasurable future income, about which each of us forms only vague and changeable guesses (see my 1979 paper). The "Pigou effect" is "explained" to undergraduates even though its currency base (Don

Patinkin, 1969, p. 1154) is a microscopic fraction of national wealth, and even though the effect is theoretically superfluous (see my 1982 paper). Many other such cases could be mentioned.

Let me speculate why pedagogic failures litter our texts. The suppliers prefer to sell thick (expensive) books. Many of the intermediaries attract and impress students by endowing our field with more scientific content than it actually has. And, the captive demanders are not given protection by rigorous and extensive journal reviews of what should be the cornerstone of economics, the textbooks. Thus they are forced—to use Frank Knight's felicitous phrase—to "know too many things that just ain't so."

### REFERENCES

- Chang, Winston W., Hamberg, Daniel and Hirata, Junichi, "Liquidity Preference as a Behavior Toward Risk is a Demand for Short-Term Securities—Not Money," *American Economic Review*, June 1983, 73, 420–27.
- Patinkin, Don, "Money and Wealth: A Review Article," *Journal of Economic Literature*, December 1969, 7, 1140–60.
- Pesek, Boris P., "Monetary Theory in the Post-Robertson 'Alice in Wonderland' Era," *Journal of Economic Literature*, September 1976, 14, 859–61.
- , "A Note on the Theory of Permanent Income," *Journal of Post Keynesian Economics*, Summer 1979, 1, 64–69.
- , "In Defense of Neoclassical Monetary Theory," *Quarterly Review of Economics and Business*, Summer 1982, 22, 126–33.

\*Department of Economics, University of Wisconsin, Milwaukee, WI 53201.

# Experimental Economics: Comment

By RONALD A. HEINER\*

In two important studies, Charles Plott (1982) and Vernon Smith (1982) assess the current state of the literature about laboratory experiments in economics. As a profession, we are becoming aware that experimental methods can be applied to our models, with cautious but growing confidence that these procedures can help us evaluate alternative theories. These are significant developments whose potential ramifications are only beginning to be explored.

Given the importance of laboratory testing, I would like to discuss a key feature of past experiments, one that has not been fully appreciated because of its central role in standard economic theory. In particular, these experiments depend on inducing agents to respond according to a prespecified value structure. This usually amounts to starting with a known and fully determinate set of demand and supply value schedules for all transacting agents. Smith, for example, specifies four major principles about how preferences are to be experimentally induced (non-satiation, saliency, dominance, and privacy; see pp. 931–35).

## I

I do not intend any criticism about induced preferences being somehow “artificial” or not “real.” Smith rightly points out that many naturally occurring or field markets also attempt to induce values by linking otherwise intangible assets (for example, airline travel vouchers) to specified bundles of rights to other more tangible commodities. Both field-induced and laboratory-induced preferences thus merit the same methodological status for theory evaluation. Plott also emphasizes that laboratory markets (although simpler and specialized compared to

natural markets) are still no less real than naturally occurring markets. They still represent legitimate special cases for which theories of general validity should be expected to apply.

The issue I raise is whether agents in non-experimental markets behave according to well-defined, complete preference orderings. What if agents must deal with uncertainty about which options are more or less preferred? Do natural or field markets simply reflect agent valuations that exist independently of exchange opportunities? Or do they help produce viable exchange patterns in the absence of fully determinate values of participating agents? Do markets simply combine pre-existing values, or do they organize a sequence of bids and offers as a feedback process helping agents cope with uncertainty in their own valuations?

Standard choice theory ignores these questions by postulating a complete preference ordering for each transactor. This applies whether preferences describe apple vs. orange consumption decisions, or an agent’s attitude toward risk between probabilistic contingencies. Yet, in each case, preference uncertainty is a plausible possibility.

For example, suppose we lined up before the customers of a grocery or department store a thousand randomly selected commodity baskets from the store. Is it likely that any of them would have secure beliefs about the relative value of all of the baskets? Moreover, if valuations for these commodity combinations are not well-defined, how it is that market prices are determined when there are no pre-existing value schedules to guide supply and demand decisions?<sup>1</sup> This does

\*Member 1984–85, The Institute for Advanced Study, Princeton, NJ 08540, and Brigham Young University.

<sup>1</sup>These issues are even more important for expected utility models of behavior under uncertainty, where agents can infer reliable probabilities of potential situations, or at least organize their experience into subjective probabilities of future contingencies. In addition, agents know all possible events that might eventuate, and all possible actions that might be useful to select. Such



not mean that agents' supply-demand behavior in field markets will fail to respond to price offers of other agents, or that a determinate market-clearing price will not be generated. Rather, the source of these regularities would have to be found in something other than agents reacting with no uncertainty about what is more preferred.

Preference uncertainty may also affect the convergence properties reported by Smith, and the efficiency measurements highlighted by Plott. The effect of changing exchange institutions or property rights may also depend on the type of uncertainty affecting agent valuations. Past experiments represent a special class of situations where agents have access to completely reliable value information to guide their interactions with each other.<sup>2</sup> But will convergence or efficiency properties be sensitive to the reliability of information that guides these interactions?<sup>3</sup>

Past experiments are also typically structured around an auction market process where agents follow specified rules to both make and respond to exchange offers of others. However, their success in dealing with each other may be significantly affected by how markets are organized, especially if agents are not guided by pre-existing valuations that are independent of how different kinds of markets generate price information.

Auction markets, for example, often produce a volatile sequence of price offers that agents must plan for and coordinate with

their individual consumption and investment decisions. Could this extra complexity contribute to uncertainty in agents' valuations, thereby conditioning where "flex" vs. "fix" price markets will evolve in an economic system? Could other features of intra- and intermarket organization (such as the evolution of money, firm organization, or ownership rules) indirectly result from how they affect agents' ability to cope with uncertain values?

The above queries about value uncertainty can be criticized in a number of ways. The simplest is to point out that these issues are outside the well-established theoretical purview of economics, which has always proceeded by postulating well-defined preferences for all transactors (whether they are commodity preferences, risk preferences, subjective probabilities, asset valuations, etc.). Past experiments are motivated within this theoretical context, and can be viewed as trying to determine the effects of this postulate for different institutional and property right settings.

Nevertheless, there is a crucial perspective from which questions of value uncertainty must be admitted, as suggested by a dominant theme of Smith's paper:

... The roots of our discipline require a complete reexamination; ... Above all, we need to develop a body of knowledge which clarifies the difference between what we have created (theory as hypothesis) and what we have discovered (hypothesis that, to date, is or is not falsified by observation). [p. 952]

---

ability to comprehend the future is much more difficult than avoiding computational mistakes in a world of known utility information over a fixed set of options (such as with a fixed product inventory from a store). See John Hey (1979, pp. 232–34; and 1981).

<sup>2</sup>This does *not* mean that these situations are irrelevant for testing. As emphasized by Plott, theories purporting general validity must therefore apply to special case laboratory environments.

<sup>3</sup>The idea of preferences as a potentially vulnerable information source is novel from the perspective of standard economic theory, and briefly discussed below in remark 3. For recent analysis about the viability of agents allowing their preferences to evolve rather than taking them as given, see Michael Cohen and Robert Axelrod (1984). See also Paul Slovic and Sarah Lichtenstein (1983, pp. 599–600, 602–03).

The assumption of well-defined preferences is the foundation from which our formal models are built; thereby representing the most important created element of our theoretical analysis. Thus, it is vital in the early stages of laboratory testing that we do not inadvertently establish a methodological precedent requiring future experiments to be structured around this created element. Rather, much in the spirit of these early experiments, we must broaden their design to investigate the effects of relaxing this created feature. For example, how are in-

dividual behavior and market structure affected by value uncertainty within different exchange situations?

A few brief remarks are directed toward this question:

1) How do we model the impact uncertainty on behavior, especially when agents cannot discern reliable probability information about the value of consequences resulting from their exchange commitments? Our formal models have almost universally ignored this issue.<sup>4</sup> I have elsewhere (1983, 1985) suggested a different theory in which genuine uncertainty, far from being unanalyzable, instead tends to produce regularity in behavior. The reason is that uncertainty requires an agent's flexibility to use information must be constrained to simpler behavior patterns that can be reliably administered. For example, a clear prediction of an inverse reaction to higher prices (i.e., the law of demand), is implied from preference uncertainty, which is not derivable if fully determinate preferences are assumed.<sup>5</sup>

2) Using uncertainty to regulate behavior has already been implicitly incorporated in the privacy condition discussed by Smith (see pp. 934-35). Maintaining privacy means agents are only informed of their own value schedules but not those of other agents. This condition makes interpersonal utility information highly uncertain, so that agents cannot reliably use such information to pursue interpersonal utility goals (i.e., they cannot reliably deviate from their pre-assigned value schedules). Consequently, agent behavior is simplified to patterns predicted from their individual value schedules, even though they might otherwise respond to interpersonal utility information.<sup>6</sup>

<sup>4</sup>Rational expectation theorists, such as Robert Lucas, even claim that *no* economic reasoning will be of value without well-defined, objective frequency information to guide behavior. See Lucas (1981, p. 224) and Steven Sheffrin (1983, p. 13).

<sup>5</sup>That is, the law of demand is implied without qualification for income effects as in standard Slutsky equation treatments. See my 1983 article, pp. 579-80.

<sup>6</sup>Similar results also apply to experiments reported by Plott that varied what agents were permitted to know about other agents. Plott states, for example, in the "incomplete information" experiments of Fouraker and Siegel, a simpler pattern of convergence around the

3) A recent example of experiments where induced preferences have an element of uncertainty is Plott and Shyam Sunder (1982). Each participant is assigned a value function that depends on realized dividend rates, which in turn depend on (two or three) states of nature (generated by a bingo device before each trading period starts). Such experiments are a legitimate but limited step to incorporating preference uncertainty. They are limited because: uncertainty is produced by a statistical process which is exogenous to the resulting trading interactions between participants; only two or three exogenous states of nature are involved; each state of nature produces the same type of mathematically prespecified change in agents' assigned value schedules (whose functional forms are also known and invariant to realized states of nature).

More basic features to address in experimental design are value uncertainties that cannot be resolved except through agents committing themselves to particular exchange offers as prompted by actual participation with other competing agents. A closely related question is how such uncertainty affects behavior within a given market when that market is itself embedded within several interconnected markets which differ in their internal organization. This may be a clue to why differently organized markets evolve into stable patterns of interdependence with each other.

4) In standard choice theory, preferences are the ultimate reference point from which agents evaluate their behavior. Thus, preferences are necessarily that which they should maximize. When values are uncertain, however, preferences do not necessarily have special status over other information that might affect behavior. Rather, they are simply another information source which may or may not reliably direct how to act.<sup>7</sup> An agent's

competitive equilibrium price was observed compared to that generated under "full information" conditions (see p. 1513).

<sup>7</sup>The same issue applies to any behavioral feedback process affected by an agent's perceptual mechanisms, such as expectations about future price or wage movements, asset yields, government policy, etc. See also the references cited in fn. 3 above.

preferences instead represent an internally generated feedback process that helps guide how to interact with the environment.

Exchange environments also enable agents to interact with each other in an organized and often repeated fashion. Markets thus provide agents with additional feedback that may help guide their behavior.

The objective now becomes to understand how different environmental settings (for example, different institutions or property rights) affect the reliability of individual and market guidance processes, thereby affecting the exchange patterns and market organization that will arise in these situations. In pursuing this objective, the analysis of feedback mechanisms studied in cybernetics may be a fruitful tool, one that has been little used in economics.<sup>8</sup>

## II

Laboratory experiments in economics have been structured around the assumption of fully determinate preference orderings, usually in the form of prespecified demand and supply value schedules. This assumption is basic to standard choice theory, but it is nevertheless a theoretically created rather than an empirically discovered feature.

Thus, if our objective in the development of laboratory testing and methodology is to distinguish between creation and discovery, a broadening of experimental design to allow for value uncertainty is necessary. In so doing, significant further discovery may be identified, especially in more complex exchange environments where the effects of

uncertainty will become increasingly important.

## REFERENCES

- Ashby, W. Ross, *An Introduction to Cybernetics*, New York: Wiley & Sons, 1963.
- Cohen, Michael D. and Axelrod, Robert, "Coping with Complexity: The Adaptive Value of Changing Utility," *American Economic Review*, March 1984, 74, 30-42.
- Heiner, Ronald A., "The Origin of Predictable Behavior," *American Economic Review*, September 1983, 73, 560-95.
- , "Uncertainty, Signal Detection Experiments, and Modeling Behavior," in R. Langlois, ed., *The New Institutional Economics*, New York: Cambridge University Press, 1985.
- Hey, John D., *Uncertainty in Microeconomics*, New York: New York University Press, 1979.
- , "Are Optimal Search Rules Reasonable?," *Journal of Economic Behavior And Organization*, 1981, 2, 47-70.
- Lucas, Robert E., *Studies in Business Cycle Theory*. Cambridge: MIT Press, 1981.
- Plott, Charles, R., "Industrial Organization Theory and Experimental Economics," *Journal of Economic Literature*, December 1982, 20, 1485-527.
- and Sunder, Shyam, "Efficiency of Experimental Security Markets with Insider Information: An Application of Rational Expectation's Models," *Journal of Political Economy*, August 1982, 90, 663-98.
- Roth, Alvin E., "Toward a Theory of Bargaining: An Experimental Study in Economics," *Science*, May 13, 1983, 220, 687-91.
- Sheffrin, Steven M., *Rational Expectations*. Cambridge: Cambridge University Press, 1983.
- Slovic, Paul, and Lichtenstein, Sarah, "Preference Reversals: A Broader Perspective," *American Economic Review*, September 1983, 83, 596-605.
- Smith, Vernon L., "Microeconomic Systems as an Experimental Science," *American Economic Review*, December 1982, 72, 923-56.

<sup>8</sup>As just one possibility, we can use information theory to measure the complexity of behavior necessary to fully maximize different preference relations (for an introductory discussion, see Ross Ashby, 1963, ch. 11). This required complexity (in order to maximize) may far exceed that which an agent can reliably administer (see my earlier paper, 1983), thus producing behavior which deviates substantially from that implied by standard optimizing models, or which may produce regularities in the way preferences evolve that are able to affect behavior.

## Experimental Economics: Comment

By DANIEL FRIEDMAN\*

In his comment on experimental methodology in this issue, Ronald Heiner reminds us that the evidence for a bedrock assumption of economic theory—the existence of well-defined individual preferences—is not very compelling in nonexperimental market settings. Of course, this point in itself is not new: it is a central theme in the first three books reviewed in the General Economic Theory section of the June 1983 issue of the *Journal of Economic Literature*, for example. Heiner goes on to suggest that the inducement of value itself be studied experimentally, and I agree that this is a potentially important area of research.

To a certain extent, “preference [or value] uncertainty” has already been incorporated in some laboratory experiments. For example, Charles Plott and Shyam Sunder (1982) and my paper with Glenn Harrison and Jon Salmon (1984) induced state-contingent preferences, the actual state being unknown during some trading periods. Well-defined *probabilities* over states were induced in these experiments however, and the results were surprisingly (to me at least) consistent with expected utility theory. It might be interesting to see what happens when such probabilities are *not* induced. For instance, subjects might be told that the state of nature has

already been determined in some unspecified manner, but will only be revealed after the market closes.

I suspect that Heiner would not be fully satisfied with such experiments. He raises the interesting possibility that individual preferences may *mutually* interact with the exchange process and market organization in nonexperimental settings. In particular, the standard assumption that preferences are independent of market activities may blind us to important phenomena. Experimentalists, I am sure, would welcome suggestions for laboratory tests of this possibility, and Heiner may find (as several other theorists already have found in different contexts) that his own alternative theory will be sharpened if he tries to devise such tests.

### REFERENCES

- Friedman, Daniel, Harrison, Glenn W. and Salmon, Jon W., “The Informational Efficiency of Experimental Asset Markets,” *Journal of Political Economy*, June 1984, 92, 349–408.
- Heiner, Ronald A., “Experimental Economics: Comment,” *American Economic Review*, March 1985, 75, 260–63.
- Plott, Charles and Sunder, Shyam, “Efficiency of Experimental Securities Markets with Insider Information,” *Journal of Political Economy*, August 1982, 90, 663–98.

\*Department of Economics, University of California, Los Angeles, CA 90024.

# Experimental Economics: Reply

By VERNON L. SMITH\*

More than in any particular method of inquiry, I think the hallmark of science is to be found in a constructively skeptical attitude toward knowledge.<sup>1</sup> The more funda-

\*Department of Economics, University of Arizona, Tucson, AZ 85721.

<sup>1</sup>The principal contribution of Popper's falsificationist methodology is, I believe, the influential attempt to develop a formal logic of skeptical inquiry. That the attempt has failed, in the sense that it has produced no defensible codified set of procedures that yield a science of scientific method (happily it would appear that all such attempts will fail), should not detract from the disciplinary value of the falsificationist perspective in approaching scientific questions. Its value to the experimentalist is to force him to ask "How can I design an experiment with the property that the set of potentially observable outcomes can be partitioned into those that are consistent with one (or a given) theory and those that are consistent with other theory(ies) (or inconsistent with the given theory)?" That experimental life is such that his effort is about as likely to fail as to succeed by no means detracts from the value of the exercise. Its value to the theorist (if he will just forgo the career-advancing primeval incentive to publish yet another technically tractable extension of the existing theory literature) is to force him to ask "How can I model this question so as to suggest (as Martin Shubik would say) a do-able experiment, and so as to yield observable implications that do not exhaust the set of possible outcomes?" That this effort will often fail does not detract from the value of the exercise. Having said this I would not want to leave the impression that experiments that are fishing expeditions in the laboratory to see what will happen are of no value; seeing what happens can be essential in defining an analytical-empirical research program. Similarly, when a theorist builds (as Buz Brock would say) castles in the air, this is not necessarily useless, for it may lead to more operational forms of theory. We should impute some non-zero probability to the proposition that Feyerabend's "anything goes" posture is right. But at this stage I think it has become pretty obvious where our professional weaknesses are concentrated. Economists, while spouting the rhetoric (Donald McClosky, 1983) of the falsificationist, are in fact verificationist to the core. We all do it. We take a proposition, conjecture, or theory, then search for supportive historical or empirical examples. As everyone ought to know, seek and ye are likely to find, whether one is a "Keynesian" or a "supply sider." What is not sufficiently appreciated is that this verificationist grubbing is a prescientific exercise in which one asks whether there is *any* supporting evidence, and

mental are the concepts and assumptions of a science, the easier it is to take them for granted and to abandon this skepticism. In this spirit, Ronald Heiner (1985) is correct in emphasizing that the "knowledge" obtained from the study of the performance of experimental markets is only as secure as the classical preference model used to induce prespecified value structures on the agents in such markets. If the purpose of an experiment is to test a theory (for example, supply and demand), and the theory is not "falsified" by the test, this in no way supports any premise of the theory which was also a premise of the experimental design. When we falsify a theory, the implication is that one or more of its assumptions about the behavior of economic agents (maximization of expected utility, commonly shared (homogeneous) expectations, risk aversion, zero subjective costs of transacting, etc.) is in question, and the immediate task is to modify the suspected behavioral assumptions of the original theory. Other assumptions—such as that agents have well-defined preferences, or know the probability distribution from which other agent values were drawn—are not brought into question by the experiment because the experimental design reproduced (or should have) the environment posited by the theory being tested. When testing formal market theories in this way, we should always be aware of the fact that *we are studying behavior within the context of our representations of the economic environment*. If any of these representations is wrong, then our studies have only increased our self-knowledge, not our knowledge of things (natural economic processes).

If we are to increase our knowledge of things, then our ultimate aim should aspire to more than discovering that the behavioral

---

how difficult it is to find; if there is none or if it is pretty hard to uncover, it suggests abandonment in the prescientific womb.

properties of our own creations are consistent with controlled experimental evidence, although the attempt to falsify these creations may be a necessary step in acquiring the conditional knowledge that can improve our theorizing ability. This is why empirical investigations of all aspects of parallelism between laboratory and field behavior are important. Similarly, our experimental and other investigations should not be *confined* to testing formal theory (for example, nomothetic experiments) since this objective requires us to impose more structure on the free play of decision making than ultimately may be justified. Finally, our research methodology should not be too rigid in testing only the market implications of a theory (or in testing only the assumptions of a theory).

Thus, John Kagel et al. (1981) have addressed direct tests of the observable implications of preference theory. This literature reports results consistent with standard preference theory (i.e., with Hicks-Slutsky income-compensated demand theory), but also with the *ad hoc* widely assumed law of demand, which Heiner (1983) is able to deduce from his model of adaptive uncertain choice.<sup>2</sup> As noted by Kagel et al., the convergence tendency reported in auction market experi-

ments "depends critically on the fact that subjects behave in a way which is consistent with utility-maximizing principles underlying consumer demand theory and that negatively sloped demand curves have been induced in the market" (p. 13). I do not wish to suggest that these studies have put to rest the issues raised by Heiner whose emphasis is on the inadequacy of standard theory when preferences are uncertain. Indeed, both the animal and especially the human preference studies of Kagel et al. exhibit "dynamic" effects or lagged responses that are not even supposed to exist in received preference theory, and which may reflect the "insecure preference beliefs" suggested by Heiner (1983). It would seem that lagged responses are inconsistent with a cognitive, calculating interpretation of preference theory, but consistent with some sort of adaptive response interpretation. Even if preference theory accounts for many agents' stationary state choices in certain experimental situations, it tells us nothing about the *processes* that yield these "good" predictions or why some agents' behavior is not consistent with the theory. The failure of animal studies to falsify demand theory can be interpreted as lending support to Heiner's (1983) emphasis on rule-governed behavior although here the "rules" are apparently programmed into the instincts (genes), unless we are prepared to accept the proposition that species other than ours have cognitive decision-making powers.

Furthermore, the numerous direct studies of individual decision making under uncertainty, over the past 25 years (see the recent papers by David Grether, 1980; Grether and Charles Plott, 1979; and especially the survey and evaluation by Paul Slovic and Sarah Lichtenstein, 1983), suggest that our theories of decision under uncertainty are in several respects inconsistent with controlled evidence. The results of these experiments are robust under replication, and various artifactual explanations of the results (that might have rescued the theory) have been systematically eliminated. The results are not to be idly dismissed by anyone with the slightest interest in evidence. What I want to suggest is that (so far as we are able to tell) experimental methods are entirely competent

<sup>2</sup>Since it seems that no one has ever produced any rigorous evidence for the existence of a price inferior good (Sir Giffen was just speculating, and had no controls on his "experiment"), its prediction by textbook theory is a curiosity (which was recognized as such by Alfred Marshall who started it all), that should have counted against the theory, just as the failure to find a planet "Vulcan" between Mercury and the sun (there was no shortage of claimed sightings), that would account for the advance of Mercury's perihelion, ultimately counted against Newtonian theory (David DeVorkin, 1983, p. 1058). In more mature sciences this might have sparked analytical interest in producing a theory consistent with the law of demand, but having other falsifiable implications. Contrarily, in economics such a theory might even be considered unpublishable because of its "lack of generality," and the Giffen good curiosity has sparked an endless preoccupation with examples of "multiple" and "unstable" equilibria based on the Walrasian adjustment mechanism, which is itself devoid of any institutional evidence. These have been good analytical exercises, but so far as I can see all this self-knowledge is good only for teaching the defenseless and uninitiated students who will form the next generation of automata.

to examine these important issues. New theory, such as that proposed by Heiner (1983) and Soo Hong Chew (1983) (also see Don Coursey, 1982) are particularly welcome at this stage in research programs using, or directly concerned with testing, preference theory under uncertainty.

However, among those who take these experimental results as a serious challenge to existing theory, not all may interpret them in the same way. In the following I will try to state some of my interpretations, and relate them, where it seems appropriate, to Heiner's work.

1) The state of experimental research on decision under uncertainty has produced many unresolved anomalies. Experimental tests of *market* theories, which explicitly assume expected utility (or value) maximization, have *not* falsified many of these theories (for example, James Cox, Bruce Roberston and myself, 1982; Plott and Louis Wilde, 1982). Yet, as indicated above, the results of direct tests are inconsistent with the expected utility hypothesis (*EUH*). Some, but not all (for example, violation of simple dominance), of these anomalies are resolved by Chew's weighted expected utility hypothesis (*WEUH*). Although at this stage I think it would be premature to abandon *EUH*, and especially its extensions, it is not premature to work on the resolution of these anomalies. (One did not reject Newton's inverse square law of attraction because the planets failed to move in perfect ellipses, nor because of the highly inconsistent observed advance in the perihelion of Mercury.)

2) One route to such a resolution may be to recognize, and to elaborate more formally, the hypothesis that subjects are more rational (in the sense of received decision theory under uncertainty) in the context of laboratory markets, than when responding to questionnaire choices among prospects, because of Heiner's conjecture that "Exchange environments also enable agents to interact with each other in an organized and often repeated fashion. Markets thus provide agents with additional feedback that may help guide their behavior" (1985, p. 263). From the study of experimental markets, I have long thought that markets may induce

greater "rationality" in behavior because they force or promote a response to, or discovery of, opportunity cost conditions, that need not be readily forthcoming when agents merely think about the choices they make. What I have in mind may be close to Armen Alchian's (1977, pp. 27-32) imitative and trial-and-error forms of conscious adaptive behavior, except that I would deemphasize the "conscious" element.<sup>3</sup>

Different forms of market organization have been found to differ in their power to induce or extract neoclassical rational behavior. Thus the English auction is slightly more efficient (97 percent of the allocations are Pareto optimal), and prices are consistently closer to the predicted second highest value among the bidders, than is the Second price sealed-bid auction (94 percent Pareto optimal allocations) (see Cox, Mark Isaac, and myself, 1983, pp. 73-75). An explanation is simple. In the English auction it is a dominant strategy to raise the standing bid if it is less than your value, and to never raise your own bid. Behaviorally, the temptation to use this strategy is irresistible, and made transparent (without thinking) by the sequential complete bid information properties of the process. In terms of Heiner's model (1983), subjects easily (I would say at low cost)

<sup>3</sup> In writing this reply, I found myself reminiscing that Adam Smith did not begin his economic analysis, as does mainstream economics, with preferences as the primeval cause of the phenomena we study. He began with a deeply insightful *observation qua axiom*, which states that man is unique among all animal species in exhibiting "the propensity to truck, barter and exchange one thing for another." (Man is not unique in revealing preferences.) After some speculation that this might be a consequence of man's ability to use language (I would speculate that man's development of language may have been in part due to the specialization and affluence made possible by markets), Smith deduced the important result that it is this power of exchanging that gives rise to wealth creating specialization, which in turn is limited by the extent of the market. Markets thus lead agents to promote ends which are no part of their intention. Except for modern scholars such as Hayek, the idea that we are studying processes that contain major elements that are *not* consciously or cognitively purposive has been lost in the pyrotechnics of straightening out Smith's little paradoxes of value. No wonder that Kenneth Boulding will tell you flat out that Smith was the first great post-Newtonian scientist.

perceive, perhaps quite unconsciously, the opportunity cost of "nonoptimal" behavior. In the Second price sealed-bid auction it is a dominant strategy to submit a bid equal to your value. But this requires *reasoning* which is in fact *very subtle*, although, as with all puzzles, it is obvious or "trivial" once you understand it.<sup>4</sup> One must perceive that if one's own bid is the highest, the price paid is the amount of the next highest bid, and therefore one's surplus to be gained is independent of the amount bid. So the "rational" bid is to maximize your chance of winning by bidding your value. About a third of the subjects recruited out of campus classrooms to participate in a sequence of Second price auctions (with values assigned independently from a distribution to all bidders in each auction), bid "as if" they perceive the implied dominant strategy from the beginning. About one-third appear to "learn" asymptotically from their success-failure experience that this is the "best" strategy. Another third do not clearly converge to the dominant strategy—some hit it irrationally, some rarely, if ever, and some bid just below value. An examination of the bids that are less than value reveals that many occur at values so low that the prospect of having the winning bid is remote. These can be interpreted as "throw-away" bids, and if in the strict sense they are "irrational," at least they are only marginally so.<sup>5</sup>

<sup>4</sup>Richard Thaler once reported in a seminar that, informally, he had gone around polling economists as to how they would bid, after describing the Second price auction procedure. This was before William Vickrey's discussion of this auction had become so well-known. He reported that very few got it "right." Most thought one should bid at least a "little" under value. The early polling of economists on Allais, Ellsberg, Second price, and other such "paradoxes" makes it clear that economists will get it "wrong" about as often as the sophomore subject (who of course we pay a monetary reward) until he or she has had considerable time to think and analyze. Incidentally, this observation provides an answer for that somewhat mythical businessman who asks, "If you're so smart why ain't you rich?" My classmate, Otto Eckstein, didn't get rich by equating price to marginal cost.

<sup>5</sup>An outstanding young economist once asked me why I was bothering to do Second price sealed-bid auction experiments since the dominant strategy property was so trivial. I encounter statements in this spirit

3) The wide variety of different experimental studies of decision making under uncertainty, yielding results inconsistent with *EUH*, are subject to different interpretations in terms of the damage they inflict on *EUH*. I think a key element in these interpretations is what Jacob Marschak (1968) long ago called the cost of thinking, calculating, deciding, and acting, which are all part of what I have called the subjective cost of transacting (*SCT*) (see my 1982 article, p. 934 and *passim*). Of course, one could argue that *EUH* and its Chew-Machina-type extensions are on the face of it inadequate theories because they leave *SCT* out of the formal apparatus. But this is much too harsh. Considerations of *SCT* are hard to formalize within a framework as general as that attempted in *EUH* and *WEUH*, that allow the latter to be deduced as limiting cases when *SCT* goes to zero, or when outcome values get large relative to a fixed *SCT*. But the modification of standard *EUH* theory by introducing *SCT* elements in particular decision-making contexts (Sydney Siegel, 1961), or in illustrative examples (my 1982 article, p. 934), point to the untapped potential of imbedding standard theories in larger (and more "rational") frameworks. How important are the *SCT* elements in the various decision contexts which yield violations of *EUH*?

(a) I think the class of violations which are due to Kahneman-Tversky framing effects (see Slovic and Lichtenstein, and the references therein), do relatively low level damage to *EUH*. The typical case here is that the options are identical in two situations except that in one the outcomes are stated in terms of what will be lost (deaths), the other in terms of what will be gained (lives saved). It seems to me that these are like elementary optimal illusions, which the individual, at comparatively low cost, can learn to recog-

---

so regularly that it has become a permanent way of life. They reveal two things: 1) how quickly, easily and matter of factly as economists we are prepared to believe our own propaganda (theorems), and to use these beliefs to insulate ourselves from evidence; and 2) how the use of experimental methods leads one to think about the world of economic knowledge in a fundamentally different way.



nize as such (for example, the individual can be taught that the death rate is one minus the survival rate. If it is hard to teach this, as it may be, then I am wrong in interpreting it as a low-cost recognition problem). We all learn that when the sun is low, the pond in the highway ahead is just a reflection, and we do not risk a rear-end collision by jamming on the brakes. In saying this, I do *not* mean to suggest that the study of framing effects is of no interest. On the contrary, these examples show how bad we can be at intuitive problem solving, and why it is important to examine a decision from alternative perspectives. Also, these examples vary in transparency. I find the second example cited by Slovic and Lichtenstein (p. 597) to involve more *SCT* than the first. I think the equivalence of the two situations in the first example could be conveyed to the uninitiated much more easily than the second. So some of these "optical" illusions may be more costly to expose than others.<sup>6</sup> Note that this *SCT* interpretation blurs the distinction often made between positive and normative economic theory, but I have never been convinced that such a distinction was very helpful.

(b) The preference reversal examples may represent still more sophisticated "opti-

cal" illusions, and may require rather more *SCT* to yield consistency of choice. However, the fact that they seem to be moderated (although they do not disappear) when the motivation is increased (Werner Pommerhne et al., 1982) is consistent with the hypothesis that subjects do seek to increase benefits net of *SCT*, even where the latter are relatively large. A preference reverser is of course vulnerable to a con game (money pump) in which a sequence of decisions will produce a loss in assets. Will a person discover his decision inconsistencies and learn the appropriate corrections in a money pump sequence? In this context *EUH* is being tested in a more market-like framework. J. E. Berg et al. (1984) report results showing that although the frequency of preference reversals is *not* reduced, the total value (dollar magnitude) of preference reversals is reduced. Apparently, there is positive interaction between the stakes, and the money pump treatment. They also report that the preference reversal phenomena tends to decline across experiments with the same subjects.

4) The results of direct laboratory tests of *EUH* have been used to explain the apparent failure of *EUH* in insurance, securities and futures markets (see Kenneth Arrow, 1982; Heiner, 1983). Although this appears to be evidence of parallelism in behavior between laboratory and field, I think we have to be particularly careful in drawing this parallel. There is first a question of the comparability of the quality of the evidence in the two environments, and second a question of whether *EUH* is failing for the same reason in the two environments. In the laboratory experiments, the situations are carefully controlled and structured, the states of nature are well-defined, and so are the outcomes. Consequently, the results are much more clearly interpretable as inconsistent with *EUH*, even if the cause of the inconsistency is an inappropriate carryover to the laboratory of Heiner's rule-governed agent whose habits have been developed in the more unstructured uncertainty environments of the field; or if, as I have suggested, the results can be interpreted in terms of *SCT*, and *EUH* is considered to be just a limiting case of a more general economic problem.

<sup>6</sup>In studies of science learning it is found that both weak and strong learners come to their first science classes with extensive "naive" theories about how the world works. They use these naive theories to explain physical events and tend, even after instruction in the new concepts and the scientific support for them, to resort to their prior theories to solve problems that differ from the textbook examples (Lauren Resnik, 1983). Of course with *EUH* we have the difficult problem of deciding when the subject is making a "mistake" (an "optical" illusion) which she can at more-or-less cost recognize as such, and when the theory is a mistake, or not relevant to the actual problem faced by the subject (which, for example, might be better represented by *WEUH* than *EUH*). The fact that about one-third of the subjects in sequential Second price sealed-bid auctions "learn" to make dominant strategy choices has been interpreted as analogous to learning that a certain mirage is an optimal illusion and the self-interest is not served by taking such a phenomenon at its face value (Vicki Coppinger et al., 1980, p. 20). Also see Thaler (1983) for a discussion of cognitive illusions and mirages in decision making.

An example may help to clarify one type of ambiguity in interpreting field observations in terms of *EUH*. Suppose you have had a sore rib for many weeks that hasn't healed. Your doctor sends you to the lab for an x-ray to "see if it is fractured." It is not fractured, and she tells you, "Well, a hairline fracture might not show, but it doesn't much matter since the treatment is the same whether it is a fracture or a contusion." You think, "This violates the Savage axioms!" Does it? How do you (or does she) know that there are only two states of nature, rib fractured or rib bruised, given that it is sore? Could a carcinoma behind the rib make it sore? There may be hundreds of causes of sore ribs, and it may not be worth anyone's trouble to list them all, or even to invest time in thinking about any significant fraction of them. A host of past experiences may have programmed your doctor to acquire information that appears to be redundant in this particular case, and she may be unable to organize these experiences into an articulate case for her actions because such a detailed cognitive treatment of every decision is neither a necessary nor a desirable feature of her *modus operandi*.

As a second example, take the reported reluctance of people to insure against rare disasters even though, since 1969, the government has offered subsidized flood insurance rates that are *below* actuarial value. Is it a fact that this violates *EUH*? If it is a fact, then I like Heiner's explanation that there is a tradeoff between the greater setup costs (these are part of what I call *SCT*) of insuring against more events of small probability, and the expected loss from failing to insure, and it is hardly economical to insure against everything. Hence, in the field situation *EUH* is failing because it formulates the *wrong economic choice problem*, by leaving out *SCT*. But it is not clear from the evidence that this is an example of the violation of *EUH*. Arrow tells us that the reason the government offered the subsidized insurance "was to relieve the pressure for the government to offer relief when floods occurred" (p. 2). If I have built a house on a flood plane, if flood planes sometimes flood (we call them 100-year floods in Arizona), and if

it is standard political procedure for the governor to declare it a disaster area, and the federal government to respond with relief when a flood occurs, then I might not buy insurance even at rates below actuarial value. (This need not be a conscious decision with people able to state that the failure to buy insurance was due to the expectation of government relief.) Without better controls on the experimental treatment variable, I don't know how to interpret the observations. This problem does not reflect on the quality of these excellent studies cited by Arrow, but on the difficulty of doing field experiments with the most desirable controls. Similar considerations apply when asking whether interest rates or stock prices vary "too much" (see the studies by Cagan and Shiller cited in Arrow). It is unclear what is to be concluded when the falsifying "facts" are no more than "an impression which many students of these markets and practitioners in them seem to have" (Arrow, p. 4).

But Stewart's finding that unprofessional speculators lose money in grain futures is cited by Arrow as "especially surprising," and he asks "why did they enter the market at all?" (p. 3). I would suggest that they do it for the same reason that people go to Las Vegas to play roulette, buy tickets in the Arizona Lottery, and play bingo at the local church on Thursdays. I don't see any way to understand these phenomena with *EUH* (it is well known that convex, risk-prefering, utility doesn't explain repetitive small stakes wagering) nor any way to understand them with Heiner's theory. I find it necessary, if not entirely satisfactory in terms of seeking a universal theory, to accept the idea that some people just simply like to gamble (ancient hunter cultures did it) and that it has commodity value, or perhaps that some people have "pathological" expectations, whether it is roulette, grain futures or stock investment. (See my 1971 article.)<sup>7</sup>

<sup>7</sup>If in all markets with uncertainty there is a subclass of participants with these "irrational" characteristics, this lowers the insurance cost of hedging and lowers the cost of capital to firms. The gamblers lose money volun-

I have no disagreement with Heiner's critique of classical preference theory, which is among the roots that should be reexamined. However, I would register disagreement with Heiner's interpretation of "privacy" as an experimental condition. As noted in my article (1982, p. 933, fn. 13; p. 935) the purpose of privacy is to maintain control over preferences. Privacy does not deny the existence of interpersonal externalities. The latter is achieved under controlled conditions by simply inducing the appropriate interdependent preferences if that is the topic of investigation. If one wishes to study the effect of "utility information" on behavior, one publicizes information on commodity allocations, or token earnings (which is analogous to income in the field), but not subject cash payoffs since these are to induce utility on allocations and indirectly on token income. The idea is to preserve the natural uncertainty about other's subjective value of allocations, and of the exchange medium.

tarily, the economy benefits and perhaps only *EUH* suffers as a predictive theory for some types of agents. But the existence of such agents in futures, stock, and option markets will cause such markets to appear to be irrational by our definitions, whereas actually these markets may be performing with high allocative efficiency, given the environment, by taking wealth away from the gamblers and giving it to the hedgers, investors, and rational expectationists. Isn't Las Vegas an exchange market between gamblers (customers) and rational expectationists (casinos)? The question may be not "Why are certain markets inefficient?" but "What is wrong with our interpretation of markets?" An important technical difference between casinos and financial markets is that in the former the agent learns immediately the outcome of her investment. But the more variable are the prices of financial instruments, the more will they have this casino characteristic, and the more attractive they will be to this type of investor. Hence, the alleged "fact" that security prices vary "too much" may be both the effect and the cause of its appeal to these kinds of investors.

I suspect that Adam Smith would wonder why there is so much modern professional interest in the internal efficiency or "perfection" of particular markets, and so little interest in what determines the extent of markets, and how this in turn may create social gains that are more important and significant than the "imperfections" in particular markets that are suggested by our theory of "rational" preferences.

## REFERENCES

- Alchian, Armen, "Uncertainty, Evolution and Economic Theory," in *Economic Forces at Work*, Indianapolis: Liberty Press, 1977.
- Arrow, Kenneth, "Risk Perception in Psychology and Economics," *Economic Inquiry*, January 1982, 20, 1-9.
- Berg, J. E., Dickhaut, J. W. and O'Brien, J. R., "Preference Reversal and Arbitrage," in V. Smith, ed., *Research in Experimental Economics*, Vol. 3, Greenwich: JAI Press, 1984.
- Chew, Soo Hong, "A Generalization of the Quasilinear Mean with Applications to the Measurement of Income Inequality and Decision Theory Resolving the Allais Paradox," *Econometrica*, July 1983, 51, 1065-92.
- Coppinger, Vicki, Smith, Vernon and Titus, John, "Incentives and Behavior in English, Dutch and Sealed-Bid Auctions," *Economic Inquiry*, January 1980, 18, 1-22.
- Coursey, Don L., "Hierarchical Preferences and Consumer Choice," unpublished doctoral dissertation, University of Arizona, 1982.
- Cox, James, Roberson, Bruce and Smith, Vernon, "Theory and Behavior of Single Object Auctions," in V. Smith, ed., *Research in Experimental Economics*, Vol. 2, Greenwich: JAI Press, 1982.
- \_\_\_\_\_, Isaac, Mark and Smith, Vernon, "OCS Leasing and Auctions: Incentives and the Performance of Alternative Bidding Institutions," *Supreme Court Economic Review*, July 1983, 2, 43-87.
- DeVorkin, David H., "Review of N. T. Roseveare, *Mercury's Perihelion from Le Verrier to Einstein*," *Science*, March 4, 1983, 219, 1058.
- Grether, David, "Bayes Rule as a Descriptive Model: The Representativeness Heuristic," *Quarterly Journal of Economics*, November 1980, 95, 537-57.
- \_\_\_\_\_, and Plott, Charles, "Economic Theory of Choice and the Preference Reversal Phenomenon," *American Economic Review*, September 1979, 69, 623-38.
- Heiner, Ronald A., "The Origin of Predictable Behavior," *American Economic Review*, September 1983, 73, 560-95.

- \_\_\_\_\_, "Experimental Economics: Comment," *American Economic Review*, March 1985, 75, 260-63.
- Kagel et al., John H., "Demand Curves for Animal Consumers," *Quarterly Journal of Economics*, February 1981, 96, 1-16.
- Marschak, Jacob, "Economics of Inquiring, Communicating, Deciding," *American Economic Review Proceedings*, May 1968, 58, 1-18.
- McClosky, Donald N., "The Rhetoric of Economics," *Journal of Economic Literature*, June 1983, 21, 481-517.
- Plott, Charles and Wilde, Louis, "Professional Diagnosis vs. Self-Diagnosis: An Experimental Examination of Some Special Features of Market With Uncertainty," in *Research in Experimental Economics*, Vol. 2, Greenwich: JAI Press, 1982, 63-112.
- Pommerehne, Werner W., Schnieder, Frederick and Zweifel, Peter, "Economic Theory of Choice and the Preference Reversal Phenomenon: A Reexamination," *American Economic Review*, June 1982, 72, 569-74.
- Resnik, Lauren, B., "Mathematics and Science Learning: A New Conception," *Science*, April 29, 1983, 220, 477-78.
- Siegel, Sydney, "Decision Making and Learning Under Varying Conditions of Reinforcement," *Annals of the New York Academy of Science*, 1961, 89, 766-83.
- Slovic, Paul and Lichtenstein, Sarah, "Preference Reversals: A Broader Perspective," *American Economic Review*, September 1983, 73, 596-605.
- Smith, Vernon L., "Economic Theory of Wager Markets," *Western Economic Journal*, September 1971, 9, 242-55.
- \_\_\_\_\_, "Microeconomic Systems as an Experimental Science," *American Economic Review*, December 1982, 72, 923-55.
- Thaler, Richard H., "Illusions and Mirages in Public Policy," *Public Interest*, Fall 1983, 73, 60-74.

# Relative Prices, Concentration, And Money Growth: Comment

By DANIEL J. RICHARDS\*

In a recent article in this *Review* (1983), Henry Chappell and John Addison provide new evidence on the "administered"-pricing thesis that links an industry's price behavior over the business cycle to its structural features, most notably, some measure(s) of the monopoly power possessed by firms in that industry. The approach taken by Chappell and Addison (C-A) is both sensible and novel. Essentially, they regress aggregate inflation measures for different groups of manufacturing industries on a distributed lag in monetary growth. The industry groups are distinguished by their degree of concentration—low, medium, or high—and the validity of the administered-pricing thesis is then tested by comparing the time pattern of response in each group to monetary impulses. Chappell-Addison conclude that this test reveals no clear response differences between the sectors of varying concentration, and hence reject the hypothesis of concentration-related administered prices.

The purpose of this comment is to show that the C-A conclusions are overly strong. Their focus on the mean lag as the sole basis of interindustry comparisons obscures important differences in the pricing performance of the low- and high-concentrated sectors. More fundamentally, the C-A test is misspecified in that it fails to differentiate between the expected and unexpected components of money growth as suggested by Robert Barro (1977). When the analysis allows for these modifications, the C-A data reveal considerable support for the administered-pricing hypothesis.

## I. The Chappell-Addison Model

A basic prediction of the administered-pricing thesis is that prices in concentrated industries will be less responsive to demand forces than prices in more competitive sectors (Adolph Berle and Gardiner Means, 1967). Chappell-Addison test this prediction by postulating the following model of inflation for any industry or group of industries:

$$(1) \quad \dot{P}_t = a_0 + \sum_{i=1}^k a_i \cdot \dot{M}_{t-i} + e_t;$$

where  $k$  = lag length,  $\dot{P}_t$  is the estimated rate of inflation, and  $\dot{M}_t$  is the rate of monetary growth in period  $t$ , respectively.  $\dot{M}_t$  is thus the principal measure of demand pressures. The monetary aggregate used to calculate  $\dot{M}_t$  is  $M_2$ . To compute inflation rates, C-A categorize all four-digit manufacturing industries into three groups according to "the industry's four-firm concentration ratio ( $CR_4$ ) as follows: high concentration  $CR_4 > 70$ ; medium concentration  $40 < CR_4 \leq 70$ ; and low concentration  $CR_4 \leq 40$ " (p. 1124). They then construct aggregate price indices and inflation rates for each group.

Equation (1) is estimated for each of the three groups using annual time-series data for the years 1959 through 1976. Ten different specifications are used, eight of which are based on the Almon lag technique, and two of which rely on ordinary least squares (*OLS*). The administered-pricing thesis is interpreted as implying a slower response of inflation to monetary growth as concentration increases, and the estimated mean lag is used as the measure of response speed. Chappell-Addison find that regardless of specification, there is little difference in the mean lag of the different industry groups. They therefore conclude that the results do not provide much support for a theory linking "sticky" or administered prices to increasing concentration.

\*Department of Economics, Hamilton College, Clinton, NY 13323. I thank Henry Chappell and John Addison who very kindly shared their data with me and provided helpful remarks on this research. I am also grateful for useful discussions with Jeffrey Pliskin, and for comments by Derek Jones and Elizabeth Jensen on an earlier draft.

## II. Is the Mean Lag the Appropriate Basis for Interindustry Comparison?

The C-A findings raise some difficult statistical issues, most of which relate to the mean lag as the basis of comparing inter-industry price responsiveness. The major difficulty with this procedure is that it confuses the size of a response with its speed. This is because the mean lag is calculated as a weighted average of the lag terms, holding the total effect constant. Given the sum of the estimated lag coefficients, the mean lag provides information about the time distribution of the overall impact this sum implies.

Summary statistics are often helpful and in this respect the mean lag is probably no worse than many others. For the purposes of Chappell-Addison, however, the fact that the mean lag comparisons contrast the time distribution of given total effects, without comparing the size of those total effects, is a crucial failure. To take an extreme example, suppose that they had found that the estimated coefficients on lagged money growth exactly doubled as one moved from the high-concentrated sector to the medium-concentrated sector, and doubled again as one moved from that sector to the low-concentrated sector. In this case, the mean lag would be precisely the same for each group. Yet the response of inflation to money growth after one, two, or several periods would be four times as great in the low-concentrated sectors as in the high-concentrated sectors. Thus, if by speed of response we mean the amount of movement that occurs per unit of time, the mean lag may clearly suppress important information.<sup>1</sup> Such ambiguity casts serious doubt on the utility of the mean lag as the sole basis of interindustry comparisons of price responsiveness. Moreover, in the case of the C-A estimates, this problem is greatly complicated by the fact that in virtually every one of their regressions, the coefficients (either explicit or implicit) on past money growth are of both positive and nega-

TABLE 1—IMPACT OF MONEY GROWTH ON INFLATION BY SECTOR

	40 ≤		
	CR4 ≤ 40	CR4 ≤ 70	CR4 > 70
Average 1-Year Impact	0.493	0.528	0.154
Average 2-Year Impact	0.923	1.210	0.441
N-Year Impact <sup>a</sup>	1.926	1.893	1.223
Adjusted R <sup>2</sup>	0.68	0.49	0.31

Source: Taken from estimates provided by Chappell and Addison.

<sup>a</sup>Sum of the lag coefficients.

tive signs. In such cases, the very meaning of the mean lag statistic comes into question because the sign of an effect and the speed of an effect are far from the same thing.<sup>2</sup>

In light of the foregoing difficulties with the mean lag statistic, I have reviewed the C-A results and constructed some alternative measures of price responsiveness for the different industries. These are displayed in Table 1, and show for each industry the average effect of money on inflation after one or two years, and in the long run (the sum of the lag coefficients) that is implied by the C-A estimates. Attention to the one- and two-year impacts is motivated by the fact that the concern over administered prices stems largely from the debate over the ability of monetary deceleration to slow inflation quickly without imposing intolerable political costs. Given the frequency of elections, one or two years seems a reasonable time to allow for the effect of money on prices. To the extent it takes longer than this, the mone-

<sup>1</sup>This confusion could occur even when the sum of the lags is identical between sectors. Thus, in the C-A regressions, the lag structures 2 2 5 1 and 3 3 0 4 have exactly the same mean lag and lag sums.

<sup>2</sup>In personal correspondence, C-A have suggested that these difficulties could be solved by using the median lag statistic. If anything, I think this statistic would be inferior to the mean lag. In making interindustry comparisons, it suffers from the same problems as the mean lag in that it obscures important information. Moreover, in the presence of oppositely signed coefficients, the median lag is a particularly *ad hoc* number. It measures the earliest time at which half the ultimate effect is felt. But, why pick this shortest time interval? With oppositely signed coefficients there will be many points at which half the effect has occurred. To exclusively focus on the earliest of these seems highly arbitrary.

tary cure for inflation may not be politically feasible.

One clear pattern emerges from Table 1. The inflationary impact of monetary growth after one, two, or even several years is substantially less in the most highly concentrated industries than in the other two sectors. After one year, monetary deceleration would decrease inflation in those industries where  $CR4$  is below 70 percent by about three times as much as in those industries where  $CR4$  exceeds this level. After two years, the difference in impact declines a bit, yet the effect of monetary deceleration is still more than twice as great in the less-concentrated groups. Even in the long run, inflation in the highly concentrated sectors appears somewhat less responsive to monetary impulses, though the difference is much less in this case than for the one- and two-year effects. With respect to the price sensitivity of the sectors of low and medium concentration, any differences are less clear. Thus, increasing concentration does not appear to have much effect in the range of  $0 < CR4 \leq 70$ . But the results clearly indicate that raising concentration above this threshold level, does make prices more sticky.<sup>3</sup> As noted, these results are based on coefficient averages over a variety of specifications. But the same coefficient pattern appears in virtually each of the individual C-A regressions.<sup>4</sup>

Table 1 also shows the adjusted  $R^2$ s C-A obtained for each industry, again averaged over the different specifications. Note that these decline systematically as the level of

TABLE 2—INFLATIONARY IMPACT OF MONEY GROWTH  
CONSTRAINED *OLS* EQUATIONS

	$CR4 \leq 40$	$40 \leq CR \leq 70$	$CR4 > 70$
Average 1-Year Impact	0.282	0.471	-0.058
Average 2-Year Impact	0.770	0.96	0.441
Adjusted $R^2$	0.443	0.285	0.220

Source: C-A data.

concentration increases, a pattern that is again repeated in the individual regressions. (Similarly, regression standard errors increase with  $CR4$ .) There are many reasons why this might occur. But one obvious suggestion is that prices in more concentrated industries are influenced by many factors other than nominal money growth, and hence, move quite independently of that growth relative to prices in more competitive sectors.

It should be added that these findings are not substantially changed when the regressions are constrained so as to preserve money neutrality. Table 2 shows the results of reestimating the *OLS* equations when the sum of the coefficients on money growth are constrained to sum to unity. These results run parallel to the more aggregate measures shown in Table 1, and are especially similar to the unconstrained *OLS* estimates.

In short, the C-A focus on the mean lag obscures some important differences in inter-industry pricing behavior. When attention is given to the effect of monetary stimulus on prices over any particular time interval, the C-A estimates do in fact suggest considerable support for the administered-pricing thesis, at least at the highest level of concentration. Corroborating evidence for this view is provided by the  $\bar{R}^2$ s for each industrial group which suggest that the highly concentrated sectors set prices in response to a variety of nonmonetary factors not considered by firms in the less-concentrated industries. However, as the next section shows, the C-A conclusions would be open to question even without these ambiguities due to a serious flaw in their analytical framework.

<sup>3</sup> Chappell-Addison did not present any statistical tests of their mean lag comparisons. Similarly, in the results shown both here and in Table 2, I have also omitted statistical tests of differences in the coefficients of each equation. The simple fact is that the number of observations is small enough, and the problem of multicollinearity large enough, that none of the implicit coefficients on  $M_{t-i}$  are significantly different from 0. Nor do the interequation differences shown here pass the normal tests of statistical significance. Nevertheless, these differences seem both clear and highly instructive.

<sup>4</sup> It is true, of course, that over any particular time period, relative prices will change between the various sectors. But these differences in the sample-specific inflation rates should be accounted for by differences in the intercept term, not in the lag coefficients.

### III. An Alternative Specification Based on the Distinction Between Expected and Unexpected Monetary Growth

Unlike Barro, Chappell-Addison do not distinguish between unanticipated and anticipated money growth. Yet this distinction is crucial to their test of the administered-pricing thesis. Rational firms may be assumed to set their nominal prices at expected profit-maximizing levels. Under a wide class of models, especially those in the spirit of equation (1), these expectations will in turn depend on anticipated monetary expansion. If this is the case, and if anticipated money growth is the primary factor in total money growth, then using the latter variable to explain the industry's chosen path of price increases greatly confuses the issue. All firms and all sectors may exhibit similar responses to total money growth, simply because they are all reacting to the same, dominant, expected portion of that growth. In the context of the administered-pricing thesis, the real issue hinges on differences in the various sectors' separate responses to the anticipated and unanticipated components of monetary growth.

To this end, I have estimated regressions of the following general form for each of the three sectors studied by C-A:

$$(2) \quad \dot{P}_t = c_0 + \sum_{j=1}^k c_j(\dot{EM})_{t-j} + \sum_{j=1}^k d_j(\dot{UM})_{t-j} + u_t,$$

where  $\dot{EM}$  and  $\dot{UM}$  are expected and unexpected money growth, respectively. These were estimated by regressing actual  $M2$  growth on exactly the same variables as used by Barro over the years 1941 through 1976: the one-and two-period lagged money growth rates, a measure of "normal" federal expenditures, and an unemployment variable. The observations for  $\dot{EM}$  are the fitted val-

ues from this equation, and those for  $\dot{UM}$  are the residuals.<sup>5</sup>

The now relevant version of the administered-pricing thesis may be stated as follows. Inflation in the concentrated sectors will be relatively more responsive to expected money growth and less responsive to unexpected money growth than it will be in the less concentrated sectors. Competitive firms will respond to all market factors that affect the equilibrium price level. These include both  $\dot{EM}$  and  $\dot{UM}$ . Firms with monopoly power, however, will gear their pricing to  $\dot{EM}$ , and be reluctant to change their prices in response to unanticipated monetary shocks.  $\dot{EM}$  acts as a focal point about which oligopolistic competitors may easily coordinate their prices. By contrast,  $\dot{UM}$  represents the unforeseen "noise" to which oligopolists will be reluctant to respond out of the fear that this would generate too much price variability and hence reduce their ability to communicate with rivals through price signaling (Carliss Baldwin, 1983). Indeed, since all firms are assumed to have equal values of  $\dot{EM}$ , and assuming that expected money growth is neutral, the real differences between firms of varying market power should mainly lie in their response to  $\dot{UM}$ .

Table 3 parallels Table 1 and presents some summary statistics derived from estimating equation (2) with four different Almon lag specifications and two *OLS* specifications. In each of the Almon lag regressions, the polynomial degree and lag structure are specified to be the same for both  $\dot{EM}$  and  $\dot{UM}$ . The maximum number of lags for each variable was limited to four due to the small number of observations in the C-A sample. (The implausibly long lag lengths estimated by C-A—on the order of

<sup>5</sup>The estimated equation is

$$\begin{aligned} \dot{EM} = & 0.0935 + 0.202 \cdot \dot{M}_{-1} + 0.453 \cdot \dot{M}_{-2} \\ & + 0.0928 \cdot FEDV + 0.0291 \cdot UN_{-1}, \end{aligned}$$

where  $FEDV$  is normal federal expenditures as defined in Barro, and  $UN$  is  $\ln[U/(1-U)]$ ;  $R^2 = 0.838$ ,  $D-W = 2.196$ .



TABLE 3—INFLATIONARY IMPACT OF EXPECTED ( $\bar{E}M$ ) AND UNEXPECTED ( $\bar{U}M$ ) MONEY GROWTH BY SECTOR

Specification	Impact After <sup>a</sup>	$CR4 \leq 40$		$40 < CR4 \leq 70$		$CR4 > 70$	
		$\bar{E}M$	$\bar{U}M$	$\bar{E}M$	$\bar{U}M$	$\bar{E}M$	$\bar{U}M$
2nd degree, Almon lag, lag length = 3 $\bar{R}^2$	1 year	2.330	0.432	2.107	0.554	1.826	-0.156
	2 years	2.670	0.349	2.999	0.831	3.270	-0.224
	$n$ years	1.731	0.770	1.586	1.312	2.359	-0.919
		0.559		0.464		0.217	
2nd degree, Almon lag, with end constraint, lag length = 3 $\bar{R}^2$	1 year	2.527	0.342	2.733	0.329	1.906	-0.048
	2 years	2.590	0.224	2.742	0.623	2.367	-0.311
	$n$ years	1.811	0.608	1.839	0.807	2.192	-0.559
		0.621		0.502		0.304	
3rd degree, Almon lag, lag length = 4 $\bar{R}^2$	1 year	3.308	0.246	3.201	0.311	2.965	-0.531
	2 years	1.729	0.222	2.911	0.294	4.718	-1.347
	$n$ years	1.411	1.504	1.781	1.237	3.766	-2.956
		0.527		0.385		0.175	
3rd degree, Almon lag, with end constraint, lag length = 4 $\bar{R}^2$	1 year	2.782	0.375	2.995	0.419	1.721	-0.140
	2 years	1.606	0.539	1.559	0.706	2.083	-0.391
	$n$ years	1.246	1.759	1.027	2.410	2.348	-0.756
		0.601		0.468		0.179	
OLS, lag length = 3 $\bar{R}^2$	1 year	2.330	0.413	2.107	0.554	1.826	-0.156
	2 years	2.664	0.596	2.999	0.830	3.269	-0.224
	$n$ years	1.730	0.770	1.586	1.310	2.358	-0.919
		0.559		0.464		0.217	
OLS, lag length = 4 $\bar{R}^2$	1 year	3.308	0.246	3.202	0.311	2.965	-0.531
	2 years	1.955	0.346	2.910	0.140	4.716	-1.346
	$n$ years	1.411	1.503	1.780	1.083	3.765	-2.954
		0.527		0.385		0.175	
OLS (constrained), lag length = 3 $\bar{R}^2$	1 year	1.219	0.461	1.215	0.592	-0.242	-0.067
	2 years	1.525	0.944	2.083	1.112	1.148	0.428
	$n$ years	1.000	1.660	1.000	2.013	1.000	0.766
		0.554		0.449		0.152	
OLS (constrained), lag length = 4 $\bar{R}^2$	1 year	3.000	0.298	2.615	0.410	0.887	-0.180
	2 years	1.525	0.572	1.874	0.723	1.053	0.174
	$n$ years	1.000	2.084	1.000	2.350	1.000	0.987
		0.569		0.432		0.018	

Source: C-A data.

<sup>a</sup>Lag sum.

seven years or more for each industry—raise further questions about their findings.)<sup>6</sup>

The results shown in Table 3 provide some support for the alternative model of adminis-

<sup>6</sup>In personal correspondence, C-A responded to this charge by noting that others (for example, Michael Bordo, 1980) have found similarly long lags. In my view, this does not make their findings any more plausible. Theory can and should be used to rule out some results. For example, in many of the C-A regressions for the low- and medium-concentrated industries, some of the biggest positive coefficients are for money growth eight and nine years earlier, accounting for one-quarter to one-half the long-run effect, and larger than the impact after one or two years. I doubt any sensible model exists that would generate this pattern.

tered prices. Regardless of specification, the relative inflationary impact of  $\bar{U}M$  is always least in the most highly concentrated sector. While differences in sectoral responses to  $\bar{E}M$  are harder to detect, the impact of positive monetary shocks after one, two, or several years is always positive and always greater in the low- and medium-concentrated industries than their impact in those industries where  $CR4$  exceeds 70 percent. Indeed, for the latter group of firms, positive monetary shocks exert not only a smaller impact, but an often deflationary one. It is not clear why this should be the case, but a possible explanation may lie in the regression

results which generate the  $EM$  and  $UM$  series. While the Durbin-Watson statistic for that regression does not indicate a serious autocorrelation problem, the estimated correlation of the residuals is negative ( $\rho = -0.054$ ). Hence, positive monetary shocks in one period imply negative shocks in the next period. In turn, these later expected negative shocks will reduce the subsequent true anticipated money growth below the value implied by  $EM$ . In other words, money growth in excess of that anticipated could cause tacit colluders to revise downward both their estimate of future expected growth and their optimal price increases. As the bottom portion of Table 3 shows, the foregoing conclusions regarding the relative impact of  $EM$  and  $UM$  are not substantially altered by constraining the sum of the lagged coefficients on  $EM$  to be unity.

#### IV. Summary and Conclusions

The Chappell-Addison paper represents an important contribution to the administered-pricing debate. Nevertheless, their major conclusion that the theory of concentration-based administered prices is refuted by the evidence is overly strong. It is based on a comparison of mean lag statistics which, at best, only partially captures interindustry differences in price responsiveness and, at worst, is highly misleading. Reexamination of the C-A coefficient estimates does in fact show a substantially slower response of inflation to monetary growth in those industries where  $CR4$  exceeds 70 percent relative to the response in more competitive sectors.

A further reservation regarding the C-A findings stems from their failure to distinguish between expected and unexpected money growth. In a world in which inflation is primarily generated by money growth and in which expectations of that growth are at least somewhat rational, the administered-price thesis suggests that the real differences in industries' pricing behavior will lie in their separate responses to that part of money growth that is predictable and that part that is not, especially the latter. Using the data and the approach described by Barro, this

modification provides additional support for the administered-price thesis in that inflation in the highly concentrated sectors appears relatively more responsive to expected money growth than to unforeseen monetary shocks.

Both the results presented here and those of Chappell and Addison are somewhat tentative due to the small number of observations and other data limitations. Further research on administered pricing is clearly needed. Chappell and Addison are to be congratulated for providing a new and useful framework to guide that research. It is hoped that the modifications to that framework suggested here will also be helpful.

#### REFERENCES

- Ackley, Gardiner, "Administered Prices and the Inflationary Process," *American Economic Review Proceedings*, May 1959, 49, 419-43.
- Baldwin, Carliss Y., "Administered Prices Fifty Years Later: A Comment on Gardiner C. Means: Corporate Power in the Marketplace," *Journal of Law & Economics*, June 1983, 26, 487-96.
- Barro, Robert J., "Unanticipated Money Growth and Unemployment in the United States," *American Economic Review*, March 1977, 67, 101-15.
- Berle, Adolph A. and Means, Gardiner C., *The Modern Corporation and Private Property*, 2d ed., New York: Macmillan, 1967.
- Bordo, Michael David, "The Effects of Monetary Change on Relative Commodity Prices and the Role of Long-Term Contracts," *Journal of Political Economy*, December 1980, 88, 1088-109.
- Chappell, Henry W., Jr. and Addison, John T., "Relative Prices, Concentration, and Money Growth," *American Economic Review*, December 1983, 73, 1122-26.
- Dhrymes, Phoebus J., *Distributed Lags: Problems of Estimation and Formulation*, 2d ed., Amsterdam: North-Holland, 1981.
- Means, Gardiner C., "Corporate Power in the Marketplace," *Journal of Law & Economics*, June 1983, 26, 467-85.

## Relative Prices, Concentration, and Money Growth: Reply

By HENRY W. CHAPPELL, JR. AND JOHN T. ADDISON\*

In his comment on our paper, Daniel Richards raises some interesting issues, but ones that are somewhat peripheral to the nature and scope of our original study. Richards makes two specific points of criticism. First, he notes that the mean lag is but one summary measure of the lag pattern, and that other such measures may indeed lead to alternative inferences being drawn from our data. Secondly, he conjectures that appropriate specification of the model should distinguish the effects of expected and unexpected money growth on sectoral inflation rates. In the light of these criticisms, he then reevaluates our empirical work, and concludes that there is rather more evidence in favor of the administered-pricing hypothesis than we reported.

Consider each point in turn. Richards states that "the fact that the mean lag comparisons contrast the time distribution of given total effects, without comparing the size of those total effects, is a crucial failure" (p. 274). By alternatively focusing on the coefficients for the first two lags of money growth and the sum of the coefficients on all lags, Richards finds that the prices of the high-concentration sector are the least responsive to money growth. But what is most clearly revealed by the results supplied by Richards in Table 1 is that most of our regressions do not preserve the long-run neutrality of money. This is the real lacuna of our own study. By the same token, the inferences drawn by Richards about "total effects" directly depend on this rather peculiar result. Rather than quickly conclude that the low-concentration sector is more responsive to money growth, we would instead argue that it would have been more appropriate at the outset to constrain the

lagged money coefficients to sum to unity. At our suggestion, Richards has estimated *OLS* equations in which this constraint is imposed. He does not replicate this procedure for the numerous Almon lag distributions employed in our original paper. We have. These regressions are consistent with Richards' findings as to the magnitudes of the first two coefficients. Also, the mean lags do appear to be shorter for the low concentration groups vis-à-vis the high-concentration group. Thus, while some revision of our conclusions may perhaps be in order, this is primarily due not to the deficiencies of the mean lag per se, but rather to the sensitivity of our results to the imposition of a long-run money neutrality constraint.

Even so, it should be emphasized that the evidence in favor of administered pricing remains very weak. As Richards notes, the coefficients on the  $M_{t-i}$  are never significantly different from zero, and the interequation differences also fail to achieve significance. That said, it is peculiar that, in Richards' results, as well as our own (from specifications imposing money neutrality), prices appear to be most responsive in the *medium*-concentration sector. Since most of the theoretical arguments relating market power to price adjustment would imply greater rigidity in markets where oligopolistic interdependence is most important (as opposed to either competitive or nearly monopolistic markets), this result is rather surprising. Perhaps future research (see below) will be able to explain this anomaly.

Richards also notes that the  $R^2$ s are consistently lower for the high-concentration group, from which he infers that prices in that sector are more independent of money growth. This is one possible interpretation, but there is an alternative that we find more compelling. The sectoral price indices are themselves averages of individual industry indices. In our sample, there were more industries in the low-concentration groups,

\*Department of Economics, University of South Carolina, Columbia, SC 29208. We acknowledge helpful comments from Geoffrey Wood and Michael Bordo, and the research assistance of Jane Pietrowski.

hence one would expect the averages for those groups to be less "noisy." This alone could explain the differences in  $R^2$  values.

We now turn to Richards' second point, namely that we should have distinguished between the anticipated and unanticipated components of money growth. We find this extension of our analysis novel in that it hints at the special problem of information confronting oligopolists seeking to maintain cartels in the face of changing economic conditions. It is this extension that differentiates his approach to administered pricing from ours. However, we have reservations about the empirical findings: our previous comments about money neutrality again apply—we have not reestimated these equations—while acute problems attach to the measurement of expected and unexpected money growth.<sup>1</sup> Empirically distinguishing between Richards' model of administered pricing and our own without far better data than currently exists would seem to be a nearly hopeless task.<sup>2</sup>

For the future, progress can possibly be made in recasting the administered-pricing debate, as suggested in our earlier paper. In particular, a fruitful line of inquiry would be to exploit the distinction between flexiprice and fixed price goods, first referred to by John Hicks. Such differences may or may not

be related to stylized market forms in any consistent fashion. We note that this distinction has proved fruitful in tackling a number of issues in the area of international finance.<sup>3</sup> A second issue, exploiting the anticipated/unanticipated money distinction, is whether oligopolists—who presumably wish to maintain the cartel—have stronger incentive to gather information to forecast events producing price changes than have firms in competitive industries.

<sup>3</sup>For example with respect to currency invoicing, the use of forward markets, and contract length. See Ronald McKinnon (1979) and Stephen Carse, John Williamson, and Geoffrey Wood (1980).

## REFERENCES

- Baily, Martin N., "Discussion," in *After the Phillips Curve: Persistence of High Inflation and High Unemployment*, Edgartown: Federal Reserve Board of Boston, 1978, 156–63.
- Carse, Stephen, Williamson, John and Wood, Geoffrey, *The Financing Procedures of British Foreign Trade*, New York: Cambridge University Press, 1980.
- Chappell, Henry W. and Addison John T., "Relative Prices, Concentration, and Money Growth," *American Economic Review*, December 1983, 73, 1122–26.
- McKinnon, Ronald I., *Money in International Exchange: The Convertible Currency System*, New York: Oxford University Press, 1979.
- Richards, Daniel J., "Relative Prices, Concentration, and Money Growth: Comment," *American Economic Review*, March 1985, 75, 273–78.

<sup>1</sup>The Barro method used by Richards to isolate the two components of money growth is fundamentally flawed (see, for example, Martin N. Baily, 1978). One obvious difficulty is of course that the model assumes that individuals form expectations on the basis of future information!

<sup>2</sup>Indeed, if anticipated money growth were only a function of lagged money growth rates, the models would be conceptually as well as empirically indistinguishable.

# Relative Risk Aversion in Comparative Statics: Comment

By ERIC BRIYS AND LOUIS EECKHOUDT\*

In a recent paper in this *Review*, Eliakim Katz (1983a) has convincingly argued that absolute and relative risk aversion should be measured in terms of terminal wealth and not as a function of changes in wealth (for example, profits).<sup>1</sup> By considering how a risk-averse competitive firm under price uncertainty reacts to an increase in the profits tax rate (with full loss offset), he has shown that its response depends upon the shape of both the absolute and relative risk aversion functions.

In this comment we indicate that Katz's suggestion to define utility in terms of final wealth should induce economists to pay attention to the very interesting, but so far little used, concept of partial relative risk aversion.<sup>2</sup> Indeed, as shown below, this concept is directly pertinent for the comparative statics analysis of a change in the profits tax rate. Besides, it gives some economic insight for a result that might otherwise look counterintuitive to many of us.

To the best of our knowledge, the notion of partial relative risk aversion was introduced by Carmen Menezes and David Hanson (1970),<sup>3</sup> who established the relationship between three measures of risk aversion (absolute; relative; partial relative) and the behavior of the risk premium under exogenous changes either in initial wealth or in the amount of risk (or both). More specifi-

cally, they proved that the properties of the partial relative risk-aversion function enable one to predict how the risk premium adjusts to a multiplicative transformation of the risk while initial wealth remains constant. It is immediately obvious that this notion is pertinent in Katz's problem, since for any output level, an increase in the tax rate on profits implies a multiplicative decrease in the risk faced by the firm without affecting its initial wealth. More precisely, by using a standard notation, final wealth ( $\bar{W}$ ) is equal to initial wealth ( $W_0$ ) plus after tax profits ( $\pi_t$ ), that is,

$$(1) \quad \bar{W} = W_0 + \pi_t \\ = W_0 + (1-t)(\bar{p}x - c(x) - B).$$

and, as shown by Katz,  $\text{sign}(dx^*/dt) = \text{sign}(\phi(W))$  where  $x^*$  stands for the optimal output level, and where

$$(2) \quad \phi(W) = -E[u''(W) \cdot (p - c'(x^*)) \\ \cdot (W - W_0)/(1-t)].$$

Now, using the partial relative risk-aversion function  $P(W_0, \pi_t)$ , defined by

$$(3) \quad P(W_0, \pi_t) = -\pi_t \\ \cdot u''(W_0 + \pi_t)/u'(W_0 + \pi_t),$$

it can be shown that an increasing  $P$  in terms of  $\pi_t$ , is sufficient to yield  $\phi(W)$  positive.<sup>4</sup> The technique of proof exactly parallels that used by Agnar Sandmo (1971) in a similar context, by considering first values of

\*Assistant Professor of Finance, CERAM, Sophia Antipolis/Valbonne, France, and Professor of Economics at Catholic Faculties of Mons, Belgium and Lille, France, respectively. We thank Pierre Hansen and Henri Loubergé for their help.

<sup>1</sup>As mentioned by Katz himself, this idea goes back to Kenneth Arrow (1965) and John Pratt (1964).

<sup>2</sup>As far as we know, this concept was used in comparative statics analysis by Peter Diamond and Joseph Stiglitz (1974), and by Joseph Hadar and William Russell (1978).

<sup>3</sup>In the same year, Richard Zeckhauser and Emmett Keeler defined the notion of "size-of-risk aversion" which is very close to the concept of partial relative risk aversion.

<sup>4</sup>As pointed out by Katz (1983b), one must be at least as careful in the use of  $P$  as in that of the relative risk aversion when it is defined as a function of  $\pi_t$  (instead of wealth). Indeed, when  $\pi_t$  can take negative values,  $P$  cannot be everywhere decreasing or constant in  $\pi_t$  under risk aversion.

$p$  inferior to the marginal cost evaluated at  $x^*$  (i.e.,  $c'(x^*)$ ) and then those superior to  $c'(x^*)$ . Thus under increasing partial relative risk aversion, a stronger fiscal pressure on profits induces risk-averse managers to increase output. This possibly counterintuitive result is very easily understood through the relationship between the partial relative risk aversion and the risk premium. When  $t$  increases, the amount of risk faced by the firm decreases,<sup>5</sup> so that under increasing partial relative risk aversion, the risk premium for the firm falls more than proportionately (see the main theorem in Menezes and Hanson). Consequently the firm is ready to assume more risk through a higher output level.

We should stress that the use of the concept of partial relative risk aversion does not contradict Katz's development but in fact puts it in another perspective. Indeed, defining the absolute ( $A$ ) and relative ( $R$ ) risk-aversion functions in terms of final wealth, one can show that<sup>6</sup>

$$(4) \quad \frac{dP}{d\pi_t} = \frac{dR}{d\pi_t} - W_0 \cdot \frac{dA}{d\pi_t}.$$

When  $R$  is nondecreasing and  $A$  nonincreasing (with at least one of them nonconstant),  $dP/d\pi_t$  is positive which, as shown above, is a sufficient condition to obtain a positive  $\phi$ . When both  $A$  and  $R$  are increasing (or decreasing),  $dP/d\pi_t$  is sign ambiguous and so is  $\phi$ .

Hence the conditions proposed by Katz in terms of  $A$  and  $R$  can be summarized in terms of  $P$ . Besides its convenience, the use of  $P$  has the advantage of giving a better economic insight about the behavior of the firm.

#### APPENDIX

A sufficient condition for  $\phi$  to be positive is that the partial risk-aversion function  $P$  is increasing in  $\pi_t$ .

#### PROOF:

For  $p \leq c'(x^*)$ , an increasing  $P$  implies

$$\begin{aligned} (A1) \quad & -(px^* - c(x^*) - B) \\ & \cdot (u''/u')(W_0 + px^* - c(x^*) - B) \\ & \leq -(c'(x^*) \cdot x^* - c(x^*) - B) \\ & \cdot (u''/u')(W_0 + c'(x^*)x^* - c(x^*) - B), \end{aligned}$$

and multiplying both sides by  $p - c'(x^*)$  reverses the inequality sign so that

$$\begin{aligned} (A1') \quad & -(p - c'(x^*))(px^* - c(x^*) - B) \\ & \cdot (u''/u')(W_0 + px^* - c(x^*) - B) \\ & \geq -(p - c'(x^*)) \cdot (c'(x^*) \cdot x^* - c(x^*) - B) \\ & \cdot (u''/u')(W_0 + c'(x^*)x^* - c(x^*) - B). \end{aligned}$$

For  $p \geq c'(x^*)$ , a relationship similar to (1') holds.

If we then multiply both sides by  $u'$  evaluated at  $(W_0 + px^* - c(x^*) - B)$  and if we take expectations, we get

$$\begin{aligned} (A2) \quad & E[-(p - c'(x^*))(px^* - c(x^*) - B) \\ & \cdot u''(W_0 + px^* - c(x^*) - B)] \\ & \geq -(c'(x^*)x^* - c(x^*) - B) \\ & \cdot (u''/u')(W_0 + c'(x^*)x^* - c(x^*) - B) \\ & \cdot E[u' \cdot (p - c'(x^*))]. \end{aligned}$$

As  $E[u' \cdot (p - c'(x^*))]$  is equal to zero by the first-order condition, we obtain the result that the left-hand side of (2) is nonnegative. But this expression is nothing but  $\phi(W)$  multiplied by  $(1 - t)$ .

#### REFERENCES

Arrow, Kenneth J., *Aspects of the Theory of Risk Bearing*, Helsinki: Yrjö Jahnsson

<sup>5</sup> Indeed in the presence of a higher  $t$ , the variance of after tax profits is lower.

<sup>6</sup> The proof is not given here since it can be found in the paper by Diamond and Stiglitz.

- Lectures, 1965.
- Diamond, Peter A. and Stiglitz, Joseph E., "Increases in Risk and in Risk Aversion," *Journal of Economic Theory*, July 1974, 8, 337-60.
- Hadar, Josef and Russell, William R., "Applications in Economic Theory and Analysis," in G. A. Whitmore and M. C. Findlay, eds., *Stochastic Dominance*, Lexington; Toronto: D. C. Heath, 1978.
- Katz, Eliakim, (1983a) "Relative Risk Aversion in Comparative Statics," *American Economic Review*, June 1983, 73, 452-53.
- \_\_\_\_\_, (1983b) "Partial Relative Risk Aversion: Limitation on its Use," Research Paper, Bar Ilan University, 1983.
- Menezes, Carmen F. and Hanson, David L., "On the Theory of Risk Aversion," *International Economic Review*, October 1970, 11, 481-87.
- Pratt, John W., "Risk Aversion in the Small and in the Large," *Econometrica*, January-April 1964, 32, 122-36.
- Sandmo, Agnar, "On the Theory of the Competitive Firm under Price Uncertainty," *American Economic Review*, March 1971, 61, 65-73.
- Zeckhauser, Richard and Keeler, Emmett, "Another Type of Risk Aversion," *Econometrica*, September 1970, 38, 661-65.

# Relative Risk Aversion in Comparative Statics: Comment

By JOHN D. HEY\*

In his paper in this *Review*, Eliakim Katz (1983) confuses an error of logic with a possible error of empirical description: the "error," which he blames Agnar Sandmo (1971) of initiating, and myself (1981) among others of perpetrating, is clearly a case of the latter rather than an instance of the former. Nevertheless, his note, properly interpreted, does serve the important purpose of drawing attention to a hitherto neglected feature of the relative risk-aversion index. As a consequence, the significance of Sandmo's result (which remains logically correct) might now be viewed in a slightly different light.

Katz makes two substantive points concerning the relative risk-aversion index  $R$ ; the first relating to its definition; the second to its argument. Katz notes that Sandmo takes profit to be the relevant argument, which implies the following definition (compare Katz's equation (1)):

$$(1) \quad R = -\pi u''(\pi) / u'(\pi).$$

Katz then goes on to argue that "...the very definition of  $R$  as given in (1) is incompatible with decreasing or constant relative risk aversion," the reason being, according to Katz, that the "... $R$  associated with... negative profits will be negative and hence smaller than the positive  $R$  associated with positive profits" (p. 452).

Unfortunately, this argument fails to note that  $u''$  may change sign as  $\pi$  changes sign. For example, consider the case of *constant relative risk aversion*. If (1) is integrated twice, then we get

$$(2) \quad u(\pi) \propto \begin{cases} \pi^{1-R} & \pi \geq 0 \\ -(-\pi)^{1-R} & \pi \leq 0 \end{cases}$$

if  $0 < R < 1$ . (There are corresponding forms for  $R$  equal to or greater than 1.)

Clearly the utility function as given in (2) displays constant relative risk aversion as defined by (1). But, equally clearly, the utility function displays risk aversion for  $\pi > 0$ , risk neutrality for  $\pi = 0$  and risk loving for  $\pi < 0$ . Thus,  $u''$  changes sign at  $\pi = 0$ , as suggested by my discussion above. Katz's argument is clearly wrong.

The questions remain whether (2) is an (empirically) sensible utility function and whether it correctly encapsulates the intuitive notion of constant relative risk aversion. On the first question, a glance at Daniel Kahneman and Amos Tversky (1979, especially p. 279) suggests that it is by no means implausible; as to the second, the intuitive notion that needs to be captured is the idea that the proportional risk premium is independent of the scale of the gamble (see John Pratt, 1964, Sections 11 and 12). Thus, for example, if the individual is indifferent between a gamble involving a 50-50 chance of  $kx_1$  or  $kx_3$  and a certainty of  $kx_2$  ( $x_1 < x_2 < x_3$ ) for  $k=1$ , then he should also be indifferent for  $k=2$  or  $k=3$ , or, indeed, for any value of  $k$ —whether positive or negative. But this requirement leads precisely to the utility function (2). For instance, if  $R = \frac{1}{2}$ , then the individual is indifferent between a 50-50 chance of  $\pi = 9$  or  $\pi = 16$  and a certainty of  $\pi = 12\frac{1}{4}$ ; similarly, the individual is indifferent between a 50-50 chance of  $\pi = 18$  or  $\pi = 32$  and a certainty of  $\pi = 24\frac{1}{2}$ ; but equally, the individual is indifferent between a 50-50 chance of  $\pi = -9$  or  $\pi = -16$  and a certainty of  $\pi = -12\frac{1}{4}$ . Thus, as I have already noted, the individual is risk loving for  $\pi$  negative—as Kahneman and Tversky have noted, a not uncommon phenomenon.

The result of Sandmo which Katz asserts is based on an error states that "an increase in a proportional profits tax (with full loss offset) will reduce, leave unchanged or increase the firm's output according as relative

\*Department of Economics and Related Studies, University of York, York YO1 5DD England.



risk aversion  $R$  is increasing, constant or decreasing" (p. 452). Consider the case when  $R$  is constant, and thus the relevant utility function that given in (2). The objective function,  $Eu(\pi)$ , in this case is given by

$$\begin{aligned} & \int_{p=-\infty}^{c(x)/x} \{-[px - c(x)](1-t)\}^{1-R} \\ & \quad \times f(p) dp + \int_{p=c(x)/x}^{\infty} \{[px - c(x)] \\ & \quad \times (1-t)\}^{1-R} f(p) dp \\ & = (1-t)^{1-R} \left\{ \int_{p=-\infty}^{c(x)/x} \{-[px - c(x)]\}^{1-R} f(p) dp \right. \\ & \quad \left. + \int_{p=c(x)/x}^{\infty} [px - c(x)]^{1-R} f(p) dp \right\}, \end{aligned}$$

from which it is clear that the value of  $t$  does not affect the optimal choice of  $x$ . Thus Sandmo is right. Similarly, the rest of Sandmo's result is correct.

Katz is clearly at liberty to generalize Sandmo's result, which he has done by postulating some initial wealth  $W_0$ . However, he should note that nowhere in his proofs has he imposed the condition that  $W$  ( $\equiv W_0 + \pi$ ) be positive. Thus, Katz's  $R$  could suffer from the same "problem" as Sandmo's  $R$ ; hence, it must come as something of a relief to Katz to learn that there was no problem in the first place!

However, Katz is not at liberty to suggest that certain combinations of signs on  $dA/dW$  and  $dR/dW$  are "nonfeasible" (see his ta-

ble, p. 453). Since  $R$  and  $A$  are related by  $R(W) = WA(W)$ , it follows that

$$(3) \quad dR/dW = A + WdA/dW.$$

If  $A$  and  $W$  are allowed to take positive or negative values, it is clear that any combinations of signs on  $dR/dW$  and  $dA/dW$  are possible. Even with  $W$  restricted to be positive,  $dA/dW$  can be positive while  $dR/dW$  is negative—as long as  $A$  is negative. Surely Katz does not consider that risk lovers are "nonfeasible"?

Thus, this comment shows that Katz is in error in attributing an error to Sandmo and others. But Katz does indirectly draw attention to a feature of the relative risk-aversion index, for negative values of its argument, which was not hitherto realized or appreciated.

## REFERENCES

- Hey, John D., "A Unified Theory of the Behaviour of Profit-Maximizing, Labour-Managed and Joint-Stock Firms Operating under Uncertainty," *Economic Journal*, June 1981, 91, 364-74.
- Kahneman, Daniel and Tversky, Amos, "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, March 1979, 47, 263-91.
- Katz, Eliakim, "Relative Risk Aversion in Comparative Statics," *American Economic Review*, June 1983, 73, 452-53.
- Pratt, John W., "Risk Aversion in the Small and in the Large," *Econometrica*, January-April 1964, 32, 122-36.
- Sandmo, Agnar, "On the Theory of the Competitive Firm under Price Uncertainty," *American Economic Review*, March 1971, 61, 65-73.

# Relative Risk Aversion in Comparative Statics: Reply

By ELIAKIM KATZ\*

The comment by Eric Briys and Louis Eeckhoudt represents an elegant reformulation of my earlier results and makes, in my view, a useful point. In contrast, the comment by John Hey is misleading. My 1983 paper is perfectly correct for risk-averse firms. The results by Agnar Sandmo (1971) and by Hey (1981) *explicitly* pertain to risk-averse firms. Hence the claim I make in my paper is fully valid.

In my earlier paper I showed that there has been a recurrent error in the literature dealing with risk-averse firms under uncertainty. For example, I show that Hey is wrong in making the statement that "*If the firm is risk averse, then the imposition of an ad-valorem tax on the objective function  $Y$  (or an increase in the rate of such a tax) will cause the firm to increase, keep unchanged or to decrease its value of  $X$  according as its index of relative risk aversion is increasing, constant or decreasing*" 1981, p. 370, emphasis added). Similarly Sandmo is wrong in combining the assumption that " $U''(\pi) < 0$ " (p. 66), with the statement that "increasing the tax rate will increase, leave constant or reduce output according as relative risk aversion is increasing, constant or decreasing" (p. 70).

Despite his obvious eagerness to absolve himself of having made an error, Hey (his comment resting on a utility function which requires *both* risk-loving and risk-averse behavior), does not (indeed cannot) make the above irrefutably wrong statements correct. Hence, it takes considerable chutzpah for Hey to claim that I am in error in attributing an error to Sandmo and others.<sup>1</sup>

Indeed, given the universally accepted assumption of risk aversion (which assumption was adopted by both Sandmo and Hey in the above quoted statements) both the point I make and the table I supply in my earlier paper are perfectly correct.

Other than repeating the claim that I am wrong, in error or confused, Hey attempts in his comment to find a utility function for which his results hold, to replace the concave utility function he now finds troublesome. It seems, however, that pickings in this area are rather slim. I leave it to the reader to judge the plausibility of a utility function wherein firms are *always* risk averse in the region of profits and *always* risk loving in the region of losses.

Finally, I should like to make one qualification to the solution offered in my paper to the problem of using a profits-based definition of risk aversion. It is not, as Hey suggests, that the solution fails when wealth is negative, since, in general, wealth is bounded below by zero. Rather, it transpires that for positive absolute risk aversion, it is always possible to find a *positive* wealth for which relative risk aversion *must* be increasing. To see this, note that relative risk aversion is equal to  $WA(W)$  where  $A(W)$  is absolute risk aversion. Hence, differentiating relative risk aversion with respect to  $W$  yields  $WA'(W) + A(W)$ , so that if  $W$  is sufficiently near zero, relative risk aversion must be increasing.

## REFERENCES

- Briys, Eric and Eeckhoudt, Louis, "Relative Risk Aversion in Comparative Statics: Comment," *American Economic Review*, March 1985, 75, 281-83.
- Hey, John D., "A Unified Theory of the Behaviour of Profit Maximizing, Labour-Managed and Joint-Stock Operation Under Uncertainty," *Economic Journal*, June 1981, 91, 364-74.

\*Department of Economics, University of Guelph, Guelph, Ontario, N6G 2W1 Canada.

<sup>1</sup>In contrast with Hey's response, Sandmo, in a communication to the managing editor dated September 23, 1982, of which I was sent a copy, fully accepted the point I make as valid and similarly accepted my analysis as correct.

\_\_\_\_\_, "Relative Risk Aversion in Comparative Statics: Comment," *American Economic Review*, March 1985, 75, 284-85.

Katz, Eliakim, "Relative Risk Aversion in Comparative Statics," *American Economic*

*Review*, June 1983, 73, 452-53.

Sandmo, Agnar "On the Theory of the Competitive Firm under Price Uncertainty," *American Economic Review*, March 1971, 61, 65-73.

## NOTES

The twelfth annual AEA Summer Program for Minority Students will be held June 17–August 9, 1985, at the University of Wisconsin-Madison. The program is open to black, Hispanic, and native American graduate students in economics, and offers intensive instruction in the core areas essential for graduate study. Students need not be economics majors, but must have completed one year of economic principles. Preference will be given to those who have also completed one year of calculus. Although the program is designed primarily for students completing the junior year, sophomores and seniors are also encouraged to apply. Participants will receive tuition, room and board, transportation, and a cash stipend. For further information and application materials, write to AEA Summer Program, Department of Economics, University of Wisconsin, 1180 Observatory Drive, Madison, WI 53706 (telephone 608+263-2441).

Commencing July 1, 1986, *SOCIETY*, the periodical of record in the social sciences, will shift to a rotating editorship. Applications for the stewardship of this publication are encouraged from qualified professionals from any of the major disciplines in social science. Since production, marketing, and fulfillment will continue to be performed at Transaction, geographic locale need not be a central consideration for candidates. Any individual interested in this position should file the following information by July 1, 1985: a brief narrative of accomplishments, including a statement of personal goals; professional resume of past appointments and writings; and a statement of resources that would be provided by your home university or institution in support of your performance of *SOCIETY*'s editorial functions. The new editor will have complete editorial autonomy in determining the contents of the periodical, and the opportunity to select a new editorial board. A modest honorarium in recognition of his or her activities will be available. Send your response directly to Irving Louis Horowitz, President, Transaction/*SOCIETY*, Rutgers University, New Brunswick, NJ 08903. Please, no telephone inquiries. All materials received will be kept confidential.

The Bureau of the Census in conjunction with the National Science Foundation and Yale University has developed a longitudinal database file of annual survey response data for individual manufacturing establishments. Initially, the Longitudinal Establishment Data (LED) File will contain annual data for the period 1972 through 1981. The source of the data for 1972 and 1977 is the Census of Manufacturers. The LED File will be updated regularly to contain data from those sources. The LED File makes available to the economic research community economic time-series of annual data for

manufacturing establishments in a form conducive to research and development. Considerable effort is being made to minimize the time and cost of maintaining the confidentiality of individual establishments requiring that estimation and computation using individual micro-data be conducted at the Census Bureau and the results screened to prevent disclosure. For additional information, contact James L. Monahan or J. R. Norsworthy, Center for Economic Studies, Bureau of the Census, Washington, D.C. 20233 (telephone 301 + 763-5684).

The Rockefeller Foundation offers grants of \$15,000 to \$30,000 to support their Program to Explore Long-Term Implications of Changing Gender Roles. Projects may examine factors that address the social, psychological, political, and/or economic phenomena associated with the rapidly changing status of women, or analyze ways in which policy may respond to these changes. Scholars and practitioners who have completed their training may apply. Awards cannot be given for completion of degree training. Proposals involving more than one investigator or more than one institution are welcome. Deadlines are March 15 and September 15, 1985. Address inquiries to Gender Roles Program, The Rockefeller Foundation, 1133 Avenue of the Americas, New York, NY 10036.

The third Biennial Conference on East Central Europe, Russia, and the Soviet Union will be held March 28–30, 1985, in Sarasota. It is sponsored by New College of the University of South Florida in Sarasota, Florida.

*Call for Papers:* The European Finance Association will hold its twelfth annual meeting, August 29–31, 1985, at the University of Bern. To serve as a chairperson, discussant, or to present a paper (copies of detailed abstracts must be sent by April 1), contact Professor Walter Wasserfallen, University of Bern, Volkswirtschaftliches Institut, Länggass-Strasse 8, 3012 Bern, Switzerland.

*Call for Papers:* The third annual meeting of the Association of Managerial Economists (AME) will be held in New York, December 29–30, 1985. Three sessions of competitively selected papers will be featured in conjunction with the ASSA meetings. Especially encouraged are papers integrating the theory of accounting, finance and industrial organizations. Submission deadline is June 15, 1985. Members and nonmembers are invited to submit papers and/or make program sugges-

tions to Professor Mark Hirschey, AME Program Chair, Graduate School of Business Administration, 1100 14th Street, University of Colorado, Denver, CO 80204 (telephone 303 + 623-4436).

---

*Call for Papers:* The twelfth annual conference of the European Association for Research in Industrial Economics will be held September 11-13, 1985, at the Department of Applied Economics, University of Cambridge. Sessions will include economics of industry and firm organization, industrial and competition policy, industrial and international trade, industry studies, regulation of industries and firms, as well as other areas of current interest in industrial economics. To present a paper, send three copies by April 1, 1985, to the Chairman of the Programme Committee, Dr. A. Singh, University of Cambridge, Department of Applied Economics, Sidgwick Avenue, Cambridge CB3 9DE, England.

---

*Call for Papers:* *Studies in Economic Analysis* is a biannual, student-edited journal soliciting research articles from both established economists and students. Submission fee is \$3.00 for nonsubscribers. Submit manuscripts to, or request format and style instructions from, The Editors *SEA*, Department of Economics, College of Business Administration, University of South Carolina, Columbia, SC 29208.

---

*Call for Papers:* The IREDU (Centre National de la Recherche Scientifique) is sponsoring an international conference, "Economics of Education: Tackling the New Policy Issues," to be held June 1986 in Dijon, France. The cosponsor is *The Economics of Education Review*. Send 1-2 page abstracts with a brief resume. Although papers dealing with developed countries will be considered, those on less developed countries will receive higher priority. Send abstracts no later than May 15, 1985, to Benoit Millot, IREDU, Faculte des Sciences-Mirande, 21004 Dijon, Cedex France.

---

World Congress of Public Finance of IIPF (International Institute of Public Finance) will be held in Madrid, August 26-30, 1985. The topic is "Public Finance: Issues and Prospects." The IIPF invites nonmembers to contribute. Contact Professor Horst Claus Recktenwald, Friedrich-Alexander-Universität Erlangen-Nürnberg, Lange Gasse 20, 8500 Nürnberg, Germany (telephone 0911 + 5302-200), or Professor H. M. van de Kar, Erasmus University, Rotterdam, Postbus 1738, 3000 DR Rotterdam, Netherlands.

---

A session on property rights analysis will be held at the 1985 meetings of the Law and Society Association,

June 6-9, San Diego, California. Any paper developing or applying property rights theory is suitable, but selection will favor those that link traditionally separate areas of the social sciences. Contact Professor David Schap, Department of Economics, College of the Holy Cross, Worcester, MA 01610.

---

The Association for Canadian Studies in the United States (ACSUS) will hold its eighth biennial conference on September 19-21, 1985, at the Franklin Plaza Hotel, Philadelphia, Pennsylvania. All areas of Canadian studies will be represented. For further details, contact the national office: ACSUS, One Dupont Circle, Suite 620, Washington, D.C. 20036 (telephone 202 + 887-6375).

---

To inaugurate its economics research program, GTE Laboratories will hold a symposium on industrial organization in Waltham, Massachusetts, August 15-16, 1985. (It immediately precedes the World Congress of the Econometric Society to be held in Cambridge.) Those interested in attending should contact Dr. Joseph Farrell, GTE Laboratories Inc., 40 Sylvan Road, Waltham, MA 02254.

---

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to air fare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States. To be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for application to be received in the ACLS office are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Attention: Travel Grant Program), 800 Third Avenue, New York, NY 10022, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting.

---

#### Deaths

Robert H. Burton, associate professor of economics, University of South Florida, August 15, 1984.

Beth Hayes, professor of managerial economics, Northwestern University, June 3, 1984.

Frances Wells Quantius, professor emeritus, Ohio State University, August 1, 1984.

Willard F. Williams, Texas Tech University, October 2, 1984.

George W. Zinke, professor emeritus of economics, University of Colorado-Boulder, September 8, 1984.

#### Retirements

Marshall R. Colberg, professor of economics, Florida State University, May 3, 1984.

Allen Early, associate professor and Director of the Center of Economic Education, West Texas State University, August 1984.

Walther Michael, associate professor of economics, Ohio State University, September 1, 1984.

Robert J. Murphy, associate professor of economics, University of South Florida, August 9, 1984.

#### Foreign Scholars

Jozef Gajda, Academy of Economics, Poland: research fellow, University of Pittsburgh, September 1984.

Klaus D. Grimm, IMS, University of Sussex and London School of Economics: Operations Research Analyst, Military Traffic Management Command, Oakland, California, October 1, 1984.

G. Shantakumar, National University of Singapore: research fellow, University of Pittsburgh, September 1984.

Takeshi Suzuki, Hosei University, Tokyo: research fellow, University of Pittsburgh, September 1984.

Katarzyna Zukrowska, Polish Institute of International Affairs: research fellow, University of Pittsburgh, September 1984.

#### Promotions

Paul Bennett: research officer and senior economist, Research and Statistics Function, Federal Reserve Bank of New York, August 1, 1984.

Wayne Carroll: assistant professor, Ohio State University, January 1, 1984.

Garey Durden: professor of economics, Appalachian State University, July 1, 1984.

Yiu-Kwan Fan: professor of economics, University of Wisconsin-Stevens Point, August 1984.

William Guthrie: associate professor of economics, Appalachian State University, July 1, 1984.

Bruce Hamilton: professor of economics, Johns Hopkins University, July 1, 1983.

James F. Haynes: associate professor of economics, Athens State College, September 1, 1984.

R. Bryce Hool: professor of economics, State University of New York-Stony Brook, September 1, 1984.

Shafiqul Islam: chief, Industrial Economics Division, International Research Department, Federal Reserve Bank of New York, July 26, 1984.

Arnold Katz: professor of economics, University of Pittsburgh, September 1984.

Masahiro Kawai: associate professor of economics, Johns Hopkins University, July 1, 1984.

Robert T. McGee: associate professor of economics, Florida State University, August 3, 1984.

Yoshimasa Nomura: assistant professor, Ohio State University, April 1, 1984.

Jack Ochs: professor of economics, University of Pittsburgh, September 1984.

Timothy Perri: associate professor of economics, Appalachian State University, July 1, 1984.

Frank J. P. Pinto: Interregional Energy Adviser, Natural Resources and Energy Division, Department of Technical Co-operation for Development, United Nations, July 1, 1984.

Peter Skaperdas: senior economist, fiscal analysis staff, Domestic Research Department, Federal Reserve Bank of New York, May 3, 1984.

Mark Walker: professor of economics, State University of New York-Stony Brook, September 1, 1984.

#### Administrative Appointments

Edward L. Claiborn: head of economics, Virginia Military Institute, August 1984.

Thomas S. McCaleb: assistant vice president of academic affairs, Florida State University, August 3, 1984.

#### New Appointments

David M. Anderson: instructor of economics, Iowa State University, August 21, 1984.

Michael J. Applegate: visiting associate professor of economics, Iowa State University, August 1, 1984.

Charles Bates: assistant professor of economics, Johns Hopkins University, July 1, 1984.

Elaine Bennett: visiting assistant professor, University of Pittsburgh, September 1984.

Erik C. Benrud: instructor of economics, Virginia Military Institute, August 1984.

Aninda Bose, State University of New York-Stony Brook: instructor, Ohio State University, October 1, 1984.

Raymond T. Brastow, University of Washington: visiting assistant professor of economics, University of Puget Sound, September 1984.

T. J. Chen, Pennsylvania State University: assistant professor of economics, University of Mississippi, fall 1983.

Soo Hong Chew: assistant professor of economics, Johns Hopkins University, July 1, 1984.

Patricia Karr Cohen, University of North Carolina-Chapel Hill: assistant professor of economics, University of Mississippi, fall 1983.

Roger Cohen, University of North Carolina-Chapel Hill: assistant professor of economics, University of Mississippi, fall 1983.

Thomas Coleman: assistant professor of economics, State University of New York-Stony Brook, September 1, 1984.

Robert Curran: economist, Developing Economics Division, External Financing Department, Federal Reserve Bank of New York, August 15, 1984.

Larry R. Dale: associate professor of economics, West Texas State University, August 1984.

Joseph S. DeSalvo: director, Center for Economic and Management Research, University of South Florida, September 1, 1984.

Nestor Dominquez: economist, fiscal analysis staff, Domestic Research Department, Federal Reserve Bank of New York, August 20, 1984.

Ellen Evans: economist, International Financial Markets Division, Financial Markets Department, Federal Reserve Bank of New York, September 10, 1984.

James R. Follain: professor of finance, University of Illinois-Urbana, August 20, 1984.

Paul B. Ginsburg, Congressional Budget Office: senior economist, Rand Corporation, August 1984.

Jeremy Gluck: economist, International Financial Markets Division, Financial Markets Department, Federal Reserve Bank of New York, July 9, 1984.

John K. Green, East Carolina University: associate professor of accounting, Washington and Lee University, July 1, 1984.

Joseph Harrington: assistant professor of economics, Johns Hopkins University, July 1, 1984.

Leonard F. Herk: assistant professor of economics, Florida State University, August 3, 1984.

Harold Hotelling, University of Kentucky: assistant professor of economics, Oakland University, August 15, 1984.

Jerry W. Johnson: visiting professor of economics, Iowa State University, August 21, 1984.

Andrew S. Joskow, Yale University: economist, Economic Policy Office, Antitrust Division, U.S. Department of Justice, September 1984.

Nancy Kane: economist, Developing Economies Division, External Planning Department, Federal Reserve Bank of New York, September 5, 1984.

Barbara H. Kehrer, The Robert Wood Johnson Foundation: vice president, The Henry J. Kaiser Family Foundation, October 1, 1984.

Michael Kennedy, University of Texas-Austin: senior economist, Rand Corporation, January 1984.

Henry B. McFarland, U.S. International Trade Commission: economist, Economic Policy Office, Antitrust Division, U.S. Department of Justice, April 1984.

Brian H. McGavin: assistant professor of economics, Florida State University, August 3, 1984.

R. Timothy McGee: economist, Business Conditions Division, Domestic Research Department, Federal Reserve Bank of New York, May 9, 1984.

Therese McGuire: assistant professor of economics, W. Averell Harriman College for Policy Analysis and Public Management, State University of New York-Stony Brook, January 1, 1985.

John Mangel, University of Georgia: visiting assistant professor of economics, University of Mississippi, 1984-85.

Roy E. Marden: corporate economist, Philip Morris, Inc., June 1983.

Kent W. Mikkelsen, University of Michigan: economist, Economic Policy Office, Antitrust Division, U.S. Department of Justice, May 1984.

Hajime Miyazaki, Stanford University: associate professor, Ohio State University, October 1, 1984.

Peter Mueser: assistant professor of economics, Johns Hopkins University, July 1, 1983.

Carl R. Neu, First National Bank of Chicago: economist, Rand Corporation, September 1984.

Daniel Nolle: economist, Industrial Economies Division, International Research Department, Federal Reserve Bank of New York, September 10, 1984.

Esmail Noori, University of Colorado: visiting assistant professor of economics, University of Mississippi, 1984-85.

Robert Porter: associate professor of economics, State University of New York-Stony Brook, September 1, 1984.

David L. Redden: instructor of economics, Virginia Military Institute, August 1984.

Duane J. Rosa: instructor of economics, West Texas State University, August 1984.

Robert Rosenthal: professor of economics, State University of New York-Stony Brook, September 1, 1984.

Christopher Rude: economist, Industrial Economies Division, International Research Department, Federal Reserve Bank of New York, September 10, 1984.

Merrile Sing, University of Wisconsin: economist, Economic Policy Office, Antitrust Division, U.S. Department of Justice, October 1984.

Kathleen J. Stirling, University of Notre Dame: assistant professor of economics, University of Puget Sound, September 1984.

Daniel D. Tatar: instructor of economics, Virginia Military Institute, August 1984.

Kuo-Chiang John Wei, University of Illinois: assistant professor of economics and finance, University of Mississippi, fall 1983.

Edward Zakkak, Louisiana State University-Baton Rouge: visiting assistant professor of finance, University of Mississippi, 1984-85.

#### Leaves for Special Appointments

Walter Enders, Iowa State University: University of California-San Diego, August 21, 1984-May 20, 1985.

Edi Karni, Johns Hopkins University: department of economics, Tel-Aviv University, 1984-85.

James R. Prescott, Iowa State University: Federal Bank of Kansas City, August 21-December 31, 1984.

#### Resignations

Peter L. Kahn, Ohio State University, September 1984.

M. Ali Khan, Johns Hopkins University: University of Illinois-Urbana, July 1, 1984.

Frank Lysy, Johns Hopkins University: World Bank, July 1, 1984.

Charles Miller, Johns Hopkins University, July 1, 1983.

Frederick Miller, Johns Hopkins University: Braxton Associates, Boston, January 1, 1984.

Thomas A. Wolf, Ohio State University, June 1984.

## NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories (please—do not send public relation releases):

- |   |   |
|---|---|
| 1—Deaths  | 6—New Appointments                                  |
| 2—Retirements                                   | 7—Leaves for Special Appointments (NOT Sabbaticals) |
| 3—Foreign Scholars (visiting the USA or Canada) | 8—Resignations                                      |
| 4—Promotions                                    | 9—Miscellaneous                                     |
| 5—Administrative Appointments                   |   |

B. Please give the name of the individual (SMITH, Jane W.), her present place of employment or enrollment: her new title (if any), new institution and the date at which the change will occur.

C. Type each item on a separate 3×5 card.

D. The closing dates for each issue are as follows: *March*, October 15; *June*, January 15; *September*, April 15; *December*, July 15.

All items and information should be sent to the Assistant Production Editor, *American Economic Review*, Room 8279, Bunche Hall, University of California, Los Angeles, CA 90024.

---



# MACROECONOMIC CONFLICT AND SOCIAL INSTITUTIONS

**Shlomo Maital and Irwin Lipnowski, Editors**

The works included in MACROECONOMIC CONFLICT AND SOCIAL INSTITUTIONS speak with a common voice on how people, as individuals and in groups, affect the outcome of macroeconomic contests. Using game theory models the authors analyze macroeconomics through matrices relating government, business, and labor interests working toward a final goal of mutually beneficial results.

The distinguished contributors include David C. Colander, Kenneth J. Koford, Clive Bull, Andrew Schotter, Paul Davidson, Benjamin Bental, Steven Plaut, Noah M. Meltz, Jeffrey B. Miller, Jerrold E. Schneider, and the editors, Shlomo Maital and Irwin Lipnowski.

1984—248 pages—\$29.95—ISBN 0-88730-034-0

# NEOCLASSICAL POLITICAL ECONOMY

**An Analysis of Rent-Seeking  
and DUP Activities**

**David C. Colander, Editor**

*Contributors include:* Mancur Olson, Jagdish N. Bhagwati, Richard A. Brecher, T.N. Srinivasan, Douglas North, Stephen P. Magee, Warren J. Samuels, Nicholas Mercuro, Michael S. McPherson, Ronald Findlay, Stanislaw Wellisz, Harold Demsetz, Elias Dinopoulos, Frederic L. Pryor, William A. Brock, Gary M. Anderson, Robert Tollison, Kenneth Koford, James T. Bennett, Thomas J. DiLorenzo, Gordon Tullock, and David C. Colander

1984—288 pages—\$32.00—ISBN 0-88410-999-2  
LC 84-11124



# BALLINGER ECONOMICS

the complete catalog  
of titles

☐ YES! Please send me  
the new BALLINGER  
ECONOMICS CATALOG  
today —

Please send to:

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
zip \_\_\_\_\_

send your request to:  
**BALLINGER  
PUBLISHING COMPANY  
ECONOMICS CATALOG**  
54 Church Street  
Cambridge, MA 02138

AAER 3/85

---

INTERNATIONAL MONETARY FUND

# STAFF PAPERS

Published quarterly in March, June, September, and December, *Staff Papers* contains studies prepared by members of the Fund's staff on monetary and financial problems affecting the world economy.

Articles appearing in the December 1984 issue are:

*Domestic Credit and Exchange Rates in Developing Countries: Some Policy Experiments with Korean Data* ..... by Leslie Lipschitz

*On Growth and Inflation in Developing Countries* ..... by Omotunde Johnson

*Credit and Fiscal Policies in a "Global Monetarist" Model of the Balance of Payments* ....  
by Peter Montiel

*Effects of Increased Market Access on Exports of Developing Countries* .....  
by Naheed Kirmani, Pierluigi Molajoni, and Thomas Mayer

*The Fund Agreement in the Courts—XX* ..... by Joseph Gold

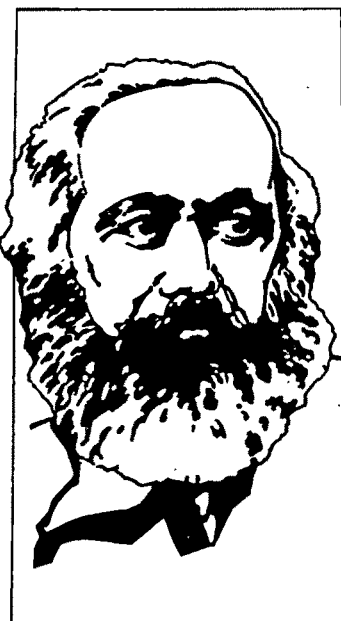
Subscriptions for *Staff Papers* (US\$15 for a volume, US\$4 for a single issue, special rates —US\$7.50 a volume—to university libraries, faculty, and students) should be sent to:

Publications Unit, Box E-189

International Monetary Fund, 700 19th Street, N.W.

Washington, D.C. 20431, U.S.A., Telephone (202) 473-7430

---



## Understanding Marx

*A Reconstruction and Critique of Capital*

Robert Paul Wolff

"This book is a *tour de force*. The author presents the Classical-Marxian theory of value and distribution in a manner completely accessible to those innocent of modern algebra, yet without any sacrifice of content. Moreover, he sets the argument in its philosophical context, thereby adding an important dimension usually missing in economic theorizing."

—Edward J. Nell, *New School for Social Research*

*Studies in Moral, Political, and Legal Philosophy*  
Marshall Cohen, Editor


246 pages. C: \$25.00. P: \$7.95

At your bookstore or

**Princeton University Press**

41 William Street  
Princeton, NJ 08540

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers



# UPDATE

## **Comparative Economic Systems Second Edition**

**Paul R. Gregory**

University of Houston, College Park

**Robert C. Stuart, Rutgers —**

The State University of New Jersey

About 450 pages • cloth • Instructor's Manual  
with Test Items • Just published

Gregory and Stuart's Second Edition offers full, balanced treatment of the theories of capitalism and socialism, illuminated with major case studies of systems in the United States, the Soviet Union, and Yugoslavia. Using a consistent framework for analysis, the authors present their material in a manner that allows students to make significant comparisons among systems.

A major revision, the Second Edition presents a thoroughly up-to-date treatment of capitalism, examines the energy crisis of the 1970s and its impact today, offers a new chapter on Yugoslavia's market socialism, and fully covers international trade. Diverse features of economic systems are clarified through studies of numerous countries, including both East and West Germany, Hungary, France, Great Britain, China, and Japan.

## **Domestic Transportation: Practice, Theory, and Policy Fifth Edition**

**Roy J. Sampson, University of Oregon**

**Martin T. Farris and David L. Shrock**

Both of Arizona State University

About 640 pages • cloth • Instructor's Manual  
Just published

A best seller in the field, *Domestic Transportation* integrates application with theory and policy, and is therefore suitable for both business administration and economics students. Updated throughout, the Fifth Edition includes new chapters on passenger transportation, international transportation, and carrier management, plus a detailed discussion of the effects of recent deregulation developments.

## **Macroeconomics**

**Norman C. Miller**

University of Pittsburgh

734 pages • Study Guide • Instructor's Manual  
1983

Miller presents all the theory, facts, empirical evidence, ideas, policy discussions, and applications necessary so students can understand economic issues and related government policy. Up-to-date, the text covers supply-side economics and international economics and is thoroughly objective in its discussions of current economic controversies.

Special pedagogical features for each chapter help students to learn and retain macroeconomic theory and apply this knowledge to current issues.

## **The Management of Financial Institutions**

**Benton E. Gup, University of Alabama**

530 pages • cloth • Instructor's Manual  
1984

Focusing upon management techniques common to the broad range of financial institutions of the 1980s, Gup's timely new text treats bank management as a useful model for other types of financial management. The practical application of theory is stressed throughout.

Chapter outlines and lists of important concepts help guide student learning.

## **Cases in Financial Management Second Edition**

**Jerry A. Viscione and George A. Aragon**

Both of Boston College

581 pages • cloth • Instructor's Manual  
1984

For adoption consideration, request examination copies from your regional Houghton Mifflin office.



## **Houghton Mifflin Company**

13400 Midway Rd., Dallas, TX 75244-5165

1900 S. Batavia Ave., Geneva, IL 60134

Pennington-Hopewell Rd., Hopewell, NJ 08525

777 California Ave., Palo Alto, CA 94304



# DEFY THE LAWS OF ECONOMETRICS.

Somewhere it is written that large-scale econometric applications require large computers. That the necessary speed and accuracy are available only on a mainframe.

It is also accepted that econometric software must be difficult to learn and complex to use. And that sizeable monthly time-sharing fees are a simple fact of life.

With all due respect to these principles of the past, we proudly present what's new: ESP™ The Econometric Software Package™ from Alpha Software®. It's a full-scale, fast and highly accurate IBM® PC and XT version of the well-known ESP mainframe program.

And it's already been selected for use by Chase Econometrics, Prudential-Bache and New York Telephone, among others.

Why? Because nothing of the mainframe program is sacrificed. You get the analytical tools you need for sophisticated modeling and forecasting. Nearly all accepted statistical and regression techniques are at your command in one truly powerful program.

In fact, ESP from Alpha is actually enhanced with features like built-in graphics and advanced data management. Plus, ESP interacts like the best personal computer software—with a spontaneous, natural feel that's instinctive, and easy to grasp. It even has on-line help and tutorials.

ESP lets you download from any major data bank including Chase and DRI. Or move your own data from other PC software like 1-2-3® Multiplan® dBASE II® and our own Data Base Manager II™.

But what's best is ESP gives you nearly instant results. Any analysis is performed before your eyes. Models, forecasts, scenarios—as many as you'd like, as often as you'd like, any time you'd like. At no incremental cost.

And you pay just \$795 for ESP from Alpha.

For more information and a demonstration version of ESP, send \$5.00 to Alpha Software Corporation, 30 B Street, Burlington, MA 01803. Or call us at 1-800-451-1018 (in Massachusetts call 1-800-462-2016).

**ESP™**  
The Econometric  
Software  
Package™



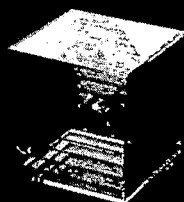
ESP and the Econometric Software Package are trademarks licensed to Alpha Software Corp. by MIKROS Corp. IBM is a registered trademark of International Business Machines Corp., 1-2-3 of Lotus, Multiplan of Microsoft, and dBASE II of Ashton-Tate. Data Base Manager II is a trademark of Alpha Software Corp. Alpha Software Corp. is a registered trademark. © 1984 Alpha Software Corp.

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

# HAVE YOU EXAMINED THE THIRD EDITION OF BAUMOL & BLINDER?

## ECONOMICS

Principles and Policy · Third Edition



# IT'S ON YOUR SHELF.

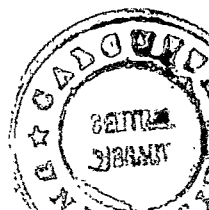
William J. Baumol and Alan S. Blinder

If you don't have an examination copy of the just-published Third Edition of **ECONOMICS: PRINCIPLES AND POLICY**, please allow us to RUSH a copy to you. Simply contact your local HBJ sales representative or your regional office, or write:

**HBJ** **HARCOURT BRACE JOVANOVIICH, PUBLISHERS**  
College Promotion,  
1250 Sixth Avenue, San Diego, CA 92101

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

v



# ALLEN & UNWIN

## APPRAISAL AND CRITICISM IN ECONOMICS

A Book of Readings  
Bruce Caldwell, Editor

A major new collection of writings by economists and philosophers on the methodology of economics. The first part contains classic articles from the positivist era from Hutchinson, Machlup, Friedman and Samuelson. Part Two covers the wide range of post-positivist contributions, including Blaug, Boland, McCloskey and Caldwell. *Paper \$12.95 Cloth \$29.50 490pp. 1984*

## KEYNESIANISM VS. MONETARISM

And Other Essays in Financial History  
Charles P. Kindleberger

"Charles Kindleberger — one of those rare beings who can move with equal ease in high theory or low history." *Barry Supple, Times Literary Supplement*

This volume brings together Professor Kindleberger's characteristically readable and stimulating essays on a wide range of topics and ideas from the financial history of the last 200 years. *Cloth \$27.50 200pp. Summer 1985*

## METHODOLOGY FOR A NEW MICROECONOMICS:

The Critical Foundations  
Lawrence A. Boland

Boland critically examines existing neoclassical paradigms of equilibria and concludes that so far no one has provided a satisfactory methodological individualistic explanation of market stability. His important new work focuses on the inherent methodological issues which must be overcome in developing innovative approaches to neoclassical economics. *Cloth \$28.50 224pp. Summer 1985*

## THE ECONOMIC ANALYSIS OF TRADE UNIONS

New Approaches and Evidence  
Barry T. Hirsch and John T. Addison

This new text reviews the burgeoning literature on labor unions and examines the effect unions had on relative wages, earnings distribution, productivity, inflation and politics. Hirsch and Addison view unions as determined by past and present worker and industry characteristics, the legal environment and other factors. *Paper \$10.95 Cloth \$27.50 256pp. 1985*

## SILICON LANDSCAPES

Peter Hall and Ann Markusen, Editors

Experts from the U.S. and the U.K. focus on the new regional development success stories — Silicon Valley, Boston's Route 128, and Britain's M4 Corridor. Their studies examine the ingredients for their success, the nature and demands of "High-Tech" industries, and their effect on local economies. *Paper \$9.95 Cloth \$25.00 160pp. February 1985*

Fifty Cross Street, Winchester, MA 01890

For MC/VISA Orders Call Toll Free  
1-800-547-8889

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

# **ECONOMICS**

**Martin Bronfenbrenner**  
*Aoyama Gakuin University, Japan*

**Werner Sichel** and **Wayland Gardner**  
*Both of Western Michigan University*

Complete hardcover edition  
Two-volume paperback edition:

**MACROECONOMICS • MICROECONOMICS**

Study Guide by Rose Pfefferbaum,  
Mesa Community College

Instructor's Manual • Test Bank • Computerized  
Test Bank • Color Transparencies • Computer  
Graphics Package • 1984

For adoption consideration, request an  
examination package from your regional  
Houghton Mifflin office.



**Houghton Mifflin**

13400 Midway Rd., Dallas, TX 75244-5165  
1900 So. Batavia Ave., Geneva, IL 60134

Pennington-Hopewell Rd., Hopewell, NJ 08525  
777 California Ave., Palo Alto, CA 94304

# **BRONFENBRENNER SICHEL & GARDNER THE RATIONAL CHOICE**

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

# AEA sponsored Group Life Insurance for you and your family— at attractive rates!

The AEA Group Life Insurance Plan can help provide valuable supplementary protection—at attractive rates—for eligible members and their dependents.

Because AEA participates in a large Insurance Trust which includes other scientific and technical organizations, the low cost may be even further reduced by premium credits. In the past eight years, insured members received credits on their April 1 semiannual payment notices averaging over 44% of their annual premium contributions. (These credits are based on the amount paid during the previous policy year ending September 30.) Of course future premium credits, and their amounts, cannot be promised or guaranteed.

Now may be a good time for you to re-evaluate your present coverage and look into AEA Life Insurance. Just fill out and return the coupon for more details at no obligation.

<b>Administrator, AEA Group Insurance Program</b> <b>1255 23rd Street, N.W.</b> <b>Washington, D.C. 20037</b>	<b>G-1</b>
Please send me more information about the AEA Life Insurance Plan.	
Name _____	Age _____
Address _____	
City _____	State _____ Zip _____

Or—call today Toll-Free 800-424-9883  
(Washington, DC area, call 296-8030)

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*



# JOURNAL OF MONEY, CREDIT, AND BANKING

## Data Storage and Evaluation Project

The *Journal of Money, Credit, and Banking* Data Storage and Evaluation Project, which is supported by National Science Foundation grant number SES-8112800, provides for the collection of data sets and programs used in selected empirical articles appearing in the journal. About two-thirds of *JMCB* articles are empirical. The aim of the project is to evaluate findings through replication of data from original sources and replication of empirical results. Data sets are available to researchers at marginal cost.

Data sets from empirical articles published in 1983 and subsequently are available as well as data from selected articles published in earlier years. A complete list of data sets now available can be obtained from the *JMCB* editorial office.

Send inquiries to: Editorial Office, *Journal of Money, Credit, and Banking*, Department of Economics, Ohio State University, 1775 College Road, Columbus, Ohio 43210-1309. Telephone: 614-422-7834.

Professional study-vacation to the

**SOVIET UNION**

for

**ECONOMISTS**

**July 6-27, 1985**

Moscow, Leningrad, Baku, Kiev, Volgograd, Repino

Hosted by Economics Section of the  
Soviet All-Union Central Council of Trade Unions

Special program includes visits such as GOSPLAN, oil fields, collective farm, research institute, factory, economic commission of city government. Opportunity to meet counterparts in various sectors of Soviet society.

**\$1895\***, all-inclusive from New York

For more information, call or write:

**Counterpart Tours**

250 W. 57th St., Suite 1428, New York, NY 10107  
1-800-223-1336 (NYS residents call (212) 245-7501)

\*May be tax deductible

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

*The real world  
comes to the classroom . . .*

# **Robert J. Barro**

*University of Rochester*

## **MACROECONOMICS**

Barro's *Macroeconomics* is the first modern presentation of the market clearing approach to macroeconomics. It provides an honest and realistic alternative to the cumbersome and unrewarding IS/LM approach of current intermediate texts.

"Robert Barro has been a leading contributor to recent developments in the analysis of the overall behavior of the economy. He has now written a lucid, comprehensive, and authoritative exposition of the current state of our understanding of these phenomena. His text covers both the latest theoretical developments and the latest empirical evidence. It can be highly recommended on both levels."

—Milton Friedman,  
Senior Research Fellow  
Hoover Institution on War, Revolution and Peace  
Stanford University

" . . . What Barro does is not only build a macroeconomic model from the ground up but also one that is meant to explain observed economic phenomena . . . if only other textbook authors would try harder to do this, students would learn more and teachers would find their job much easier."

—James Barth,  
The George Washington University

1984    580 pp.

You owe it to yourself and your students to examine this important new text. To request a complimentary copy, write to Lisa Berger, Dept. 5-1926. Please include course name, enrollment, and title of present text.

**JOHN WILEY & SONS, Inc.**  
**605 Third Avenue, New York, N.Y. 10158**

In Canada 22 Worcester Road, Rexdale, Ontario M9W 1L1



5-1926

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

*Keeping pace with the latest  
developments in economics*

---

**THE THEORY AND  
PRACTICE OF ECONOMETRICS**  
SECOND EDITION

**George G. Judge**, *University of Illinois*

**William E. Griffiths**, *University of New England*

**R. Carter Hill**, *University of Georgia*

**Helmut Lütkepohl**, *Universität Osnabrück*

**Tsoungh-Chao Lee**, *University of Connecticut*

Now the most complete treatment of major econometric problems has been revised and updated. The authors present the most systematic and up-to-date view of econometric problems, their statistical consequences, remedies, alternatives, and future research. Includes new chapters on asymptotic distribution theory, Bayesian inference, time series, and simultaneous equation statistical models. The distribution lag chapters have been rewritten to tie-in with time-series chapters.

Due January 1985 approx 1050 pp. ISBN 0-471-89530-X

---

*Other New Titles:*

**AGRICULTURAL ECONOMICS AND AGRIBUSINESS**, Third Edition

**Gail L. Cramer**, *Montana State University*

**Clarence W. Jensen**, *Montana State University*

January 1985 approx 475 pp. ISBN 0-471-87871-5

**ECONOMICS OF PRODUCTION**

**Bruce R. Beattie**, *Montana State University*

**C. Robert Taylor**, *Montana State University*

January 1985 approx 320 pp. ISBN 0-471-80810-5

---

*1984 Titles of Interest:*

**AMERICAN MONEY AND BANKING**

**Maxwell J. Fry**, *University of California*

**Raburn M. Williams**, *University of Hawaii*

**MONETARY AND FINANCIAL ECONOMICS**

**James L. Pierce**, *University of California, Berkeley*

**URBAN LAND ECONOMICS**

**Michael A. Goldberg**, *University of British Columbia*

**CASES IN MANAGERIAL ECONOMICS**, Second Edition

**Bernard J. Winger**, *University of Dayton*

To request a complimentary copy, write to Lisa Berger, Dept. 5-1926. Please include course name, enrollment, and title of present text.

**JOHN WILEY & SONS, Inc.**

605 Third Avenue, New York, N.Y. 10158

In Canada 22 Worcester Road, Rexdale, Ontario M9W 1L1



5-1926

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

## **Books that matter are Basic.**

### **Thinking Economically**

**How Economic Principles Can Contribute to Clear Thinking**

**MAURICE LEVI**

In this witty and beautifully written book, Maurice Levi shows us with a wealth of interesting real-life examples exactly how economists approach problems and how "thinking economically" can help us make better decisions. Economic thinking, according to Levi, need not be—and indeed should not be—restricted to abstract problems or government level policy making. "A valuable and clearly written book that shows the great benefits of using economics in day-to-day thinking."—JACOB A. FRENKEL, University of Chicago.

\$16.95

### **Principles of Money, Banking, and Financial Markets**

**LAWRENCE S. RITTER & WILLIAM L. SILBER**

Ritter and Silber has been widely praised as the most complete, literate, and scholarly introduction to the field of money, banking, and financial markets available. Guided by 200 questionnaire responses from instructors, the Fourth Edition has been thoroughly revised and reorganized to incorporate the most up-to-date material. Augmented by extensive illustrative material, the book provides full, balanced coverage of all the topics dealt with in standard money and banking courses. Text, \$25.95 Instructor's Manual, gratis Study Guide, \$10.00

### **The Ideal Worlds of Economics**

**BENJAMIN WARD**

Is there only one "true" way of understanding the economic universe in which we live? In this book, a distinguished economist demonstrates that there are in fact *three* distinct ways of describing how economies work—a liberal way, a radical way, and a conservative way—each consistent with presently available evidence and each in harmony with widely held moral views. Indeed, he presents these "optimal economic world views" so persuasively that the reader will find himself fully convinced by each—until he comes to the next.

Now available in its entirety, paper, \$18.00

### **Losing Ground**

**American Social Policy 1950–1980**

**CHARLES MURRAY**

"The Administration's new 'bible.'"—LEONARD SILK, *The New York Times*. "A great book. . . . The first important 'third generation' critique of recent social policy."—*The Wall Street Journal*. "The great value of *Losing Ground* is that, without bile and without rhetoric, it lays out a stark truth that must be faced: two decades of well-meaning programs to erase racism and poverty in the U.S. have left those at the very bottom of the ladder worse off than ever."—*Business Week*.

\$23.95

**Forthcoming**

**Beyond Human Scale**

**ELI GINZBERG & GEORGE VOJTA**

**April**

**Choosing Elites**

**ROBERT KLITGAARD**

**April**

**Basic Books, Inc.**

**10 E. 53rd St., New York, NY 10022**

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

## Misregulating Television

*Network Dominance and the FCC*

**Stanley M. Besen,  
Thomas G. Krattenmaker,  
A. Richard Metzger, Jr.,  
and John R. Woodbury**

This book provides evidence that it is the FCC regulations — rather than the networks — that have been principally responsible for shaping and limiting the viewing options of the public. The authors argue that, by standards of competition, programming diversity, and localism, the restraints that the FCC has tried to impose have been either ineffective or harmful.

**Cloth \$24.00 208 pages**

## The Soul of Modern Economic Man

*Ideas of Self-Interest,  
Thomas Hobbes to Adam Smith*  
**Milton L. Myers**

"Milton Myers deserves congratulations for producing this concise, insightful, and highly readable account of the intellectual origins of capitalist economics." — *The Annals of the American Academy of Political and Social Science*

**Paper \$6.95 158 pages**

## Macroeconomics and Micropolitics

*The Electoral Effects of Economic Issues*  
**D. Roderick Kiewiet**

Kiewiet systematically investigates the influence of economic concerns — particularly inflation and unemployment — on voting decisions in national elections.

"This is an important book. . . . Scholars and budding scholars in political science, economics, and sociology will find it indispensable." — Richard Brody, Stanford University

**Paper \$5.00 186 pages**

## New Studies

*In Philosophy, Politics,  
Economics, and the History of Ideas*  
**Friedrich A. Hayek**

This selection of Hayek's recent essays and lectures includes several important ones that have never before been published in English. The essays cover a wide range of subjects but are united by a common philosophical approach. Included are Hayek's contributions to the debate on inflation, a historical and systematic account of liberalism, and the Nobel Memorial Lecture delivered in Stockholm in 1974, when Hayek was awarded the Nobel Prize in Economic Science.

**Paper \$12.50 322 pages**

## Anticipations of the General Theory? And Other Essays on Keynes

**Don Patinkin**

Patinkin examines the much-debated question of whether Maynard Keynes' greatest work — *The General Theory of Employment, Interest, and Money* — was an instance of simultaneous discovery.

"This is a book to enjoy — there are no loose ends, no undocumented details, no casual references. The research is precise, detailed, confident and competent."

— *Wall Street Review of Books*

**Paper \$10.00 304 pages**

## Selfishness, Altruism, and Rationality

*A Theory of Social Choice*  
**Howard Margolis**

"Howard Margolis's intriguing ideas . . . provide an alternative to the crude models of rational choice that have dominated economics and political science for too long."

— *Times Literary Supplement*

**Paper \$8.00 206 pages**

The University of **CHICAGO** Press

5801 South Ellis Avenue, Chicago, IL 60637

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

## *Economics from MIT*

### **Market Structure and Foreign Trade:**

Increasing Returns, Imperfect Competition, and the International Economy

*Elhanan Helpman and Paul R. Krugman*

Relating current theoretical work to the main body of trade theory, this book offers entirely new material on contestable markets, oligopolies, welfare, and multinational corporations, and offers new insights on external economies, intermediate inputs, and trade composition. \$22.50

### **Exchange Rate Management under Uncertainty**

*edited by Jagdeep S. Bhandari*

*Foreword by Richard N. Cooper*

This collection of original work explores issues at the frontiers of exchange rate determination and open economy macroeconomics. \$30.00

### **Monetary Policy in Our Times**

*edited by Albert Ando, Hidekazu Eguchi,  
Roger Farmer, and Yoshio Suzuki*

These eight essays cover monetary policy in an uncertain world, domestic and international aspects of monetary policy, and policies to overcome stagflation. In particular, they recognize and provide a lively forum for the different views of academic and central bank economists. \$25.00

### **Equilibrium and Macroeconomics**

*Frank Hahn*

This book collects Frank Hahn's less technical essays on economic theory. With his characteristic style, wit, and principle he explores the concept of equilibrium and its "usefulness," the problematic role of money in the general equilibrium framework, and the shortcomings of monetarists, rational expectationists, and neo-Ricardians. \$29.95

*Now available in paperback*

### **Money and Inflation**

*Frank Hahn*

This rigorous essay in abstract theory is also a sustained and brilliant assault on the new monetarism. \$5.95

### **Trade Policy in the 1980s**

*edited by William R. Cline*

These twenty-two contributions by many of the top scholars in the field cover a broad range of important areas of trade conflict. "...contains ideas and data that will be relevant for some time to come."—*Foreign Affairs*  
Institute for International Economics. \$15.00

*Write for our current economics catalog*

**The MIT Press**

28 Carleton Street, Cambridge, MA 02142

## **Worker Cooperatives in America**

**Edited by ROBERT JACKALL and HENRY M. LEVIN**

This comprehensive study of a vital form of workplace democracy examines the history, dynamics, challenges, and potential of worker cooperatives in America. It illuminates key aspects of the cooperative phenomenon and lays the groundwork for future research.

\$24.95

## **The Suburban Squeeze**

**Land Conversion and Regulation in the San Francisco Bay Area**

**BY DAVID E. DOWALL**

While the tension between housing demand and land use control is felt across the U.S., the problems posed by local land use controls and state environmental regulations are most acute in California. Dowall focuses on the effects of land use and environmental regulations in San Francisco Bay Area communities, providing insights into the dynamics of land use regulation in any highly regulated market. *California Series in Real Estate Economics and Finance* and *California Series in Urban Development*. \$24.50

## **Latin Journey**

**Cuban and Mexican Immigrants in the United States**

**by ALEJANDRO PORTES and ROBERT L. BACH**

This volume details an eight-year survey of Mexican and Cuban immigrants in the 1970s. The authors describe patterns of occupational and economic development, cultural adaptation, and social relationships both within the ethnic circle and in the larger community. \$45.00 cloth, \$11.95 paperback

## **The Reconstruction of Western Europe**

**1945-1951**

**by ALAN S. MILWARD**

A comprehensive study of the economic and political reconstruction of western Europe since 1945, by one of the foremost historians on the subject. Milward offers strikingly new interpretations within a truly major work of scholarship. \$38.50

## **The Economic Rise of the Habsburg Empire**

**1750-1914**

**by DAVID F. GOOD**

Little is known about the economic history of the Habsburg Empire. In this survey of economic development in the late stages of one of Europe's former great powers, Good provides overwhelming evidence that the Habsburg Empire's economy did not fail in its last one hundred years. \$32.00

## **Chinese Business Under Socialism**

**The Politics of Domestic Commerce, 1949-1980**

**by DOROTHY J. SOLINGER**

Focusing on three determinants of commercial policy in China—the planned economy, policy conflict among the leadership, and economic underdevelopment—Solinger looks at how the Chinese have organized the balance between public and private sectors in handling the treatment of the private sector, the disposal of consumer goods, and free markets for agricultural produce. \$30.00

## **Fathers Work for Their Sons**

**Accumulation, Mobility and Class Formation**

**in an Extended Yoruba Community**

**by SARA S. BERRY**

In this original study of Yorùbá cocoa farmers and their descendants in western Nigeria, Berry examines the consequences of agricultural commercialization of economic development, political mobilization, and social change. \$24.50

## **University of California Press**

**Berkeley 94720**

**NEW IN 1985 FROM**  
**MACMILLAN**

**MONEY, BANKING,  
AND FINANCIAL MARKETS**

*Second Edition*

by ROBERT AUERBACH (University of California, Riverside)

*WITH: Instructor's Manual and Study Guide*

**BASIC ECONOMICS**

by JAMES A. DYAL and NICHOLAS KARATJAS

(both of Indiana University of Pennsylvania)

*WITH: Instructor's Manual and Study Guide*

**THE APPLIED THEORY OF PRICE**

*Second Edition*

by DONALD N. McCLOSKEY (University of Iowa)

*WITH: Instructor's Manual and Study Guide*

**MANAGERIAL ECONOMICS**

by J. WILSON MIXON (Bassett, Parks, Silberberg & Finch)

and NOEL D. URI (Federal Trade Commission and  
George Mason University)

*WITH: Instructor's Manual and Study Guide*

**INTERNATIONAL FINANCE AND  
OPEN ECONOMY  
MACROECONOMICS**

by FRANCISCO RIVERA-BATIZ (Indiana University) and

LUIS RIVERA-BATIZ (University of Chicago)

*WITH: Solutions Manual*

*For immediate attention to your textbook needs, call the Macmillan  
Mainline toll-free at 1-800-223-3215, or write:*

**MACMILLAN**  
PUBLISHING COMPANY

COLLEGE DIVISION • 866 THIRD AVENUE • NEW YORK, NY 10022

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*



**Now  
in paperback  
from  
Pantheon**

## **MINORITY REPORT**

**What's Happened to Blacks,  
Hispanics, American In-  
dians, and Other Minorities  
in the Eighties**

**Edited by  
LESLIE W. DUNBAR**

"Required, compelling reading."

—VERNON E. JORDAN, JR.

"Solid and rounded...A repository of  
information and ideas for study,  
discussion, and action."

—Kirkus Reviews

Paperbound \$8.95

## **THE NATIONS WITHIN**

**The Past and Future of  
American Indian  
Sovereignty**

**by VINE DELORIA, JR.  
and CLIFFORD LYTLE**

"This new book on Indian self-rule is  
the most informative that I have  
seen in my own half-century of  
reading."

—SOL TAX,

Professor of Anthropology,  
University of Chicago

"Superb." —Philadelphia Inquirer

Paperbound \$10.95

## **REBUILDING AMERICA**

**A Blueprint for the New  
Economy**

**by GAR ALPEROVITZ  
and JEFF FAUX**

"While the American economy is enjoying  
a recovery, there are some major cracks that  
will have to be repaired. *Rebuilding America*  
begins the process of outlining the changes  
that will be needed."

—LESTER C. THUROW

"An impressive performance...careful,  
detailed and persuasive."

—ROBERT LEKACHMAN,  
Professor of Economics,  
City University of New York

Paper \$10.95, cloth \$20



Now at your bookstore

**PANTHEON BOOKS**

201 East 50th Street, New York, N.Y. 10022

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

## **NEW FROM 21ST CENTURY PRESS**

### ***International Trade and the Future of the West***

**John M. Culbertson**

Recent events show that the shift of industries and jobs between nations depends on *absolute* cost—not *comparative advantage*. This book demonstrates that the straightforward use of economic analysis explains these events, and that the tools of economics have been widely misused in support of “free trade.”

In the radically new situation that now exists, “free trade” causes international wage-competition and the shift of industries and jobs from high-wage to low-wage nations. This process threatens an economic fall of the West. It will make for world-wide poverty in a world in which the burden of the massive and worsening overpopulation of some countries is spread through wage-competition to all.

To escape this destructive process and bring about trade that actually benefits the participating nations requires not “free trade” but trade that is managed so as to be in balance—the condition that is required to make comparative advantage applicable—and to lead to a justifiable distribution of industries among nations. This book raises fundamental issues for introductory students and students of international trade. It permits critical consideration of the relation between recent events and traditional textbook doctrines on international trade.

242 pages    hardback    1984    \$17.95 (plus \$1.50 shipping)  
paperback    1985    \$8.95; at faculty discount, \$7.15; examination copy, \$5.35  
(each plus \$1.20 shipping and handling)

### ***The Dangers of “Free Trade”***

**John M. Culbertson**

This booklet gives a brief statement of the new interpretation of international trade of *International Trade and the Future of the West*. It is suitable as a supplementary text or partial text in introductory as well as advanced courses. It will serve to open up discussion of the revolutionary changes in international trade, the way economic analysis is to be applied to them, and the effects that will follow from different kinds of trade policies.

### ***Competition, Constructive and Destructive***

**John M. Culbertson**

“Competition” is depicted by economic theory and the current deregulation doctrine as inherently beneficial—so that what the economy needs is only *more “competition.”* But experience and much of the literature of economics imply that “competition” is not necessarily beneficial, its effects depending on what firms *are competing at*. In this view, to achieve constructive competition requires judicious rules and regulations, which will prevent the destructive competition that otherwise will prevail.

This booklet will enlarge and enliven discussion in introductory and advanced courses. It provides a basis for considering controversial issues in the application of economic analysis to current cases of economic regulation and deregulation.

both about 40 pages    paperback    1985  
\$2.95; at faculty discount, \$2.35; examination copy, \$1.80  
(each plus \$.80 shipping and handling)

*Prices are subject to change. Examination copies are for consideration as a course text. For use as course texts, these books are available in text editions at lower cost. Buyers with Wisconsin address please add 5 percent of price for sales tax.*



**21ST CENTURY PRESS**  
P.O. Box 5010, MADISON, WI 53705

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

# **The Journal of International Economics and Economic Intergration Offers**

## **\$5,000**

### **For the First Annual Daeyang Prize in Economics**

- The Journal of International Economics and Economic Integration is published biannually by the Institute for International Economics, King Sejong University, Seoul, Korea.
- The purpose of the Journal of International Economics and Economic Integration is to support and encourage research in the area of international trade, international finance and other related economic issues that include general professional interest in international economic affairs.
- The Journal of International Economics and Economic Integration welcomes unsolicited manuscripts, which will be considered for publication by the Editorial Board.
- The Editorial Board will choose fourteen manuscripts for publication on an annual basis.
- The Editorial Board will choose the best manuscript out of the fourteen to be awarded the \$5,000 Daeyang Prize in Economics.
- The manuscripts, which should be accompanied by an abstract of no more than 100 words, should be typewritten, double-spaced, in English with footnotes, references, figures, tables and any other illustrative material on separate sheets.
- Three copies of the manuscript and all accompanying material should be submitted to the following address by October 31, 1985 for consideration for the 1986 publication.

**Institute for International Economics  
King Sejong University  
Seongdong-Ku, Seoul, Korea**



Economics Institute  
1030 13th Street  
Boulder, Colorado 80302 U.S.A.

**GUIDE TO GRADUATE STUDY IN ECONOMICS,  
AGRICULTURAL ECONOMICS, AND DOCTORAL  
DEGREES IN BUSINESS AND ADMINISTRATION**

in the United States of America and Canada, 7th edition, 544 pages

edited by Wyn F. Owen and Larry R. Cross

Published by the Economics Institute—a nonprofit educational corporation sponsored by the American Economic Association and endorsed by the American Agricultural Economics Association.

The **GUIDE** is an indispensable reference book for prospective graduate students—both domestic and foreign—and their advisors and sponsors. Comparative analyses of the programs are given. Over three hundred individual programs are described.

ORDER FORM

Economics Institute  
Publications Center  
1030 13th Street  
Boulder, Colorado 80302 U.S.A.

Please send me \_\_\_\_\_ copy(ies) of the **GUIDE** at \$33.00 per copy. For foreign orders, please enclose an additional \$3.00 for shipping and handling.

\_\_\_\_\_ I enclose \$\_\_\_\_\_ (check or international money order)

\_\_\_\_\_ Please bill me \$\_\_\_\_\_.

\_\_\_\_\_ Charge \$\_\_\_\_\_ to my \_\_\_\_\_ Mastercard, \_\_\_\_\_ Visa, or

\_\_\_\_\_ American Express      Number \_\_\_\_\_

Expires \_\_\_\_\_ Authorized Signature \_\_\_\_\_

For faster service on credit card orders only, call 303-492-8417 ext. 23.

Name \_\_\_\_\_ Title \_\_\_\_\_

Address \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_

Zip Code \_\_\_\_\_ Country \_\_\_\_\_  
(plus four)

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

## ECONOMETRIC SOFTWARE FROM TSP

**TSP™ 4.0** : a greatly enhanced edition of the Time Series Processor software package, in use at over 650 sites worldwide. The new release features

- Convenient data manipulation for time series and cross section data; full missing value capabilities.
- All standard econometric techniques, ordinary least squares, two and three-stage least squares, polynomial distributed lags, autoregressive correction.
- Nonlinear estimation for single and multi-equation models.
- Forecasting and solution of simultaneous models.
- Box-Jenkins identification, estimation, and forecasting.
- Databank facility for IBM, Vax, Prime, and Univac computers.
- Many other features such as matrix algebra which makes TSP a complete programming language for econometric problems.

**RATS** : a complete package for vector autoregressive models of time series, written by Litterman and Doan of VAR Econometrics. It features the estimation and simulation of VAR and ARIMA models, as well as a powerful set of spectral analysis techniques.

**PROBIT** : for the binary probit model.

**MLOGIT** : for the multinomial or conditional logit model.

**TSP International** 928 Mears Court, Stanford, CA 94305, (415) 326-1927



### KING SAUD UNIVERSITY Riyadh, Saudi Arabia

The College of Administrative Sciences, King Saud University, has vacant faculty positions (Professor, Associate Professor, Assistant Professor) for Ph.D. holders, who would be employed on contract basis as of commencement of the academic year 1985-1986, which begins on July 27, 1985.

The language of instruction at the college is ARABIC except for the Hospital Administration Program. The college has the following academic departments: (1) Accounting, (2) Business Administration, (3) Economics, (4) Law, (5) Political Science, (6) Public Administration, (7) Quantitative Methods, and (8) Hospital Administration, which is a part of the Department of Public Administration.

**Noteworthy benefits:**

- Free return air tickets annually for faculty member and his family.
- Furnished accommodation or housing and furnishing allowances.
- Monthly transport allowance.
- Relocation allowance.
- End-of-service gratuity.
- Free medical and dental care covering family.
- Contribution by University to tuition fees of non-Arabic-speaking children.

Interested academicians are kindly requested to send non-returnable photocopies of their academic diplomas and specialized experience certificates, together with their resumes (including lists of their publications and references) and written applications indicating the position applied for and the subjects the applicant is qualified to teach, to:

Dean of College  
King Saud University  
P.O. Box 2459  
Riyadh 11451, Saudi Arabia

and

Ms. Aida Ganim  
King Saud University Office  
2425 West Loop South, Suite 450  
Houston, TX 77027

Please include address and telephone number (if available) so, if selected, you may be contacted.

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

## **The Economics of Unemployment**

### ***A Comparative Analysis of Britain and the United States***

**JAMES J. HUGHES and RICHARD PERLMAN**

In this comprehensive treatment of the issues and problems of unemployment, the authors integrate discussions of measurement, theory, and policy with empirical evidence drawn from postwar experience in the United States and the United Kingdom.

Cloth \$39.50 Paper \$12.95

## **The Economist's View of the World**

### ***Governments, Markets, and Public Policy***

**STEVEN E. RHOADS**

In a clear, lively, and nontechnical style, the author explains and assesses the way in which economists view the world of public policy. The book explains why Democratic and Republican economists so often agree about micro-policy issues, and why they are frequently at odds with many politicians, consumer advocates and business and union leaders.

"I have been bowled over by the eloquence and incisiveness of the argument in the critical chapter on 'The Economist's Consumer and Individual Well Being' . . . the author shows such a clear understanding of the economist's emphasis on opportunity cost, on benefit cost comparisons, and the like."—Alfred E. Kahn, Professor of Economics, Cornell University

"I doubt that there is a better book for the typical person interested in improving his or her ability to think clearly about public policy issues."—Edgar O. Olsen, Professor of Economics, University of Virginia

Cloth about \$37.50 Paper about \$12.95

## **Energy Policy in America Since 1945**

### ***A Study of Business-Government Relations***

**RICHARD H.K. VIETOR**

Since the development of a sound energy policy for cheap fossil fuel—coal, petroleum, and gas—has become a central concern in shaping the economy, we have needed a balanced, authoritative account of the history of energy policy in the US, which, at last, we have in Richard Vietor's book.

"This book should be read by anyone who is interested in energy policy, business history, or business-government relations—which means everyone."—Robert B. Stobaugh, Director of The Energy Project, Harvard Business School

*Studies in Economic History and Policy:*

*The United States in the Twentieth Century*

\$29.95

## **Money and Markets**

### ***Essays by Robert W. Clower***

**DONALD A. WALKER, Editor**

Robert Clower's essays, brought together by Donald Walker, constitute a well-rounded treatment of the major problems in monetary economics, and the volume as a whole demonstrates how the study of monetary economics may extend current knowledge of the short-run economic fluctuations and prove useful in developing policy options to ameliorate them.

\$37.50

**Cambridge  
University  
Press**

## **Drastic Measures**

### ***A History of Wage and Price Controls in the United States***

**HUGH ROCKOFF**

A history of America's use of wage and price controls from colonial times through Richard Nixon's experiment with controls which began in 1971. Rockoff concludes that temporary controls may serve as useful "pain-killers" that permit more fundamental cures, such as restrictive monetary and fiscal policies, to be used to maximum effect.

*Studies in Economic History and Policy:  
The United States in the Twentieth Century*

\$29.95

## **Monetary Politics**

### ***The Federal Reserve and the Politics of Monetary Policy***

**JOHN T. WOOLLEY**

"... a technically proficient analysis of a highly complex area of public policy written in an uncommonly lively and stylistically attractive way. Woolley's grace as a writer makes the intricacies of monetary policy accessible to non-specialists and reflects the drama and significance inherent, but rarely revealed, in a subject usually regarded as esoteric."—*Congressional Quarterly Weekly Report*

"Woolley's study is a refreshing alternative to both the short-term preoccupations with market movements and the trivialisation of the analysis of political influences ... a scholarly study, well documented at every point. ... essential reading for anyone who wants to understand monetary policy making in the United States."—*The Banker*

\$37.50

## **Game Theoretic Analysis of Voting in Committees**

**BEZALEL PELEG**

A theoretical and completely rigorous analysis of voting in committees that provides mathematical proof of the existence of democratic voting systems which are immune to the manipulation of preferences of coalitions of voters.

*Econometric Society Monographs in Pure Theory*

\$37.50

## **Instrumental Variables**

**R.J. BOWDEN and D.A. TURKINGTON**

The authors present and develop the methodology of instrumental variables in its most general and explanatory form. They clearly illustrate these instrumental variables techniques and apply them to a range of problems, both linear and nonlinear.

*Econometric Society Monographs in Quantitative Economics*

\$34.50

## **Rivalry and Central Planning**

### ***The Socialist Calculation Debate***

**DON LAVOIE**

Professor Lavoie offers a serious challenge to conventional thinking in contemporary comparative systems and the economics of socialism. He argues that the lessons of the socialist calculation debate is that planning and markets are fundamentally alternative coordination mechanisms.

*Historical Perspectives on Modern Economics*

About \$34.50

All prices subject to change. Order from your bookstore or call our Customer Service Department at 1-800-431-1580 (outside New York State and Canada). MasterCard or Visa accepted.

**Cambridge University Press, 32 East 57th Street, New York, NY 10022**

The American Economics Association (AEA) is now soliciting applications to host the AEA Summer Program for Minority Students, for three summers beginning in 1986.

This program is now in its twelfth year and is currently at the University of Wisconsin at Madison. Previous host institutions have been Berkeley, Northwestern, and Yale.

The intent of the program is to increase the number of Blacks, Hispanics, and Native Americans pursuing the Ph.D. in Economics. In recent years the course of study has been an intensive eight to ten week program in intermediate microeconomics and macroeconomics, at the honor's level, and courses in econometrics or mathematical economics.

Funding for the program has been provided by the hosting institutions and grants to the AEA. Applications should be sent to Professor Donald J. Brown, chairman of the AEA Committee on the Status of Minority Groups in the Economics Profession, CSMGEP, no later than September 1, 1985.



# In all probability, you'll be hearing a lot about this new Wadsworth text...

**BASIC STATISTICS IN BUSINESS AND ECONOMICS, 4th Edition**  
George W. Summers, Arizona State University; William S. Peters, University of New Mexico; and Charles P. Armstrong, University of Rhode Island

**H**ere is an outstanding text by authors who, as teachers, know the difference between a basic text that is adequate and one that is outstanding. This new edition builds on the strengths of previous editions in *all* areas:

**1** *Aptness of examples and exercises.* The text's applied flavor links the realistic business situations in the text to the many examples and the carefully graded problems.

**2** *Emphasis on logical thinking.* This edition clarifies and expands coverage on hypothesis testing and two-way ANOVA, and focuses even more strongly on an intuitive approach to concepts, without resorting to rigorous mathematics.

**3** *Abundant practice materials.* Data sets and computer problems are greatly expanded in this edition, giving students ample chances to fix ideas and procedures.

**4** *A solid support system.* The Student Supplement, prepared by the authors, provides programmed learning units, diagnostic self-tests, self-correcting exercises, and achievement tests. Approximately 768 pages. 7-3/8 x 9-1/4. Casebound. October 1984. © 1985. Student Study Guide. Instructor's Manual.

## ALSO OF INTEREST

### AVAILABLE NOW

**INTERNATIONAL  
ECONOMICS, Second Edition**  
By Robert J. Carbaugh,  
University of Wisconsin/  
Eau Claire

**COMPARATIVE ECONOMIC  
SYSTEMS, Second Edition**  
By John E. Elliott, University  
of Southern California

**THE ECONOMICS OF  
DEVELOPING COUNTRIES**  
By E. Wayne Nafziger,  
Kansas State University

For additional information about these  
and other Wadsworth texts, write  
Stephanie Surfus, Economics and Deci-  
sion Sciences Editor

**W WADSWORTH  
PUBLISHING  
COMPANY**

Ten Davis Drive  
Belmont, CA 94002

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

# Authoritative answers to critical problems in economics...

## DEBT TRAP

Rethinking the Logic of Development

Richard W. Lombardi

An incisive look at banking, the third world debt crisis, and world stability, this new book establishes both a conceptual and practical model for re-assessing the present development logic that Lombardi states, "has reached the limits of its usefulness."

192 pp. (tent)  
\$29.95 (tent)

March 1985  
ISBN 0-03-003007-2

## PAPER PROPHETS

A Social Critique of Accounting

Tony Tinker

This highly original volume takes a hard look at the world of modern accounting, and shows how accounting policy and practice affect the quality of all our lives and the world in which we live.

224 pp.  
\$19.95 (tent)

May 1985  
ISBN 0-03-001657-6

## RECONSTRUCTING MARXIAN ECONOMICS

Marx Based Upon A Sraffian Commodity  
Theory of Value

Spencer J. Pack

Reconstructing Marxian Economics addresses the unresolved problem posed by the late economists Joan Robinson and Ian Steedman who criticized Marx's labor theory of value from a Sraffian perspective.

144 pp. (tent)  
\$28.95 (tent)

April 1985  
ISBN 0-03-003092-7

Available through your local bookseller, or order directly from:

**PRAEGER**  
PUBLISHERS

521 Fifth Avenue, New York, N.Y. 10175

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

## **GLOBAL PLAN FOR EMPLOYMENT**

A New Marshall Plan

Angelos Th. Angelopoulos

234 pp.  
\$29.95

1983  
ISBN 0-03-063798-8

## **SINO-AMERICAN ECONOMIC EXCHANGES**

The Legal Contributions

Guiguo Wang

224 pp. (tent)  
\$32.95 (tent)

January 1985  
ISBN 0-03-001659-2

## **GRASSROOTS DEVELOPMENT IN LATIN AMERICA AND THE CARIBBEAN**

Oral Histories of Social Change

Robert Wasserstrom

240 pp. (tent)  
\$32.95 (tent)  
Paper edition  
\$11.95 (tent)

February 1985  
ISBN 0-03-001689-4  
ISBN 0-03-001692-4

## **INSURANCE MARKETS**

Information, Problems, and Regulation

David A. Lereah

176 pp. (tent)  
\$29.95 (tent)

February 1985  
ISBN 0-03-001019-5

## **WORLD TRADE ISSUES**

Regime, Structure, and Policy

Young Whan Kihl and James M. Lutz

304 pp. (tent)  
\$36.95 (tent)

January 1985  
ISBN 0-03-063057-6

## **HISPANIC YOUTH**

Emerging Workers

Richard Santos

244 pp. (tent)  
\$30.95 (tent)

January 1985  
ISBN 0-03-071739-6

Available through your local bookseller, or order directly from:

**PRAEGER**  
**PUBLISHERS**

521 Fifth Avenue, New York, N.Y. 10175

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

# **Princeton**

## **What Drives Third World City Growth?**

**A Dynamic General Equilibrium Approach**

**Allen C. Kelley and Jeffrey G. Williamson**

To account for dramatic population increases in the Third World's cities since the 1950's, the authors use a Computable General Equilibrium Model to assess the importance of various factors in Third World city growth, and to discuss the impact of the OPEC-induced fuel scarcity and other changes on city growth "slowdown" in the late 1970's. LPE: \$14.50. C: \$40.00

## **The British Fertility Decline**

**Demographic Transition and the Crucible of the Industrial Revolution**

**Michael S. Teitelbaum**

"This is the first publication to marshal systematic evidence based on county level data for the whole of the British Isles for the demographic transition period. The data are entirely new and constitute a significant contribution to the field." \$40.00

—E.A. Wrigley, *London School of Economics*

*Written under the auspices of the Office of Population Research, Princeton University*

41 William Street **Princeton University Press** Princeton, NJ 08540

---

## **NOW AVAILABLE**

**Collected and on Microfiche for the first time.**

# **ADAM SMITH REFERENCES TO THE WEALTH OF NATIONS**

Pergamon Press has compiled the complete texts of monographs used by Adam Smith as references for *The Wealth of Nations* and put them on microfiche. Part of a new Pergamon series, *History of Economics*, this major collection goes beyond Edwin Cannan's classic *Index of Authorities*.

Published in two segments, and now available for immediate delivery, this comprehensive collection comprises:

- **Over 2000 Fiche**
- **Over 180,000 Filmed Pages**
- **Two Important Supplementary Collections**
  - 46 works from Adam Smith's personal library
  - 22 early editions of *The Wealth of Nations*

For further information write or call:



## **PERGAMON PRESS**

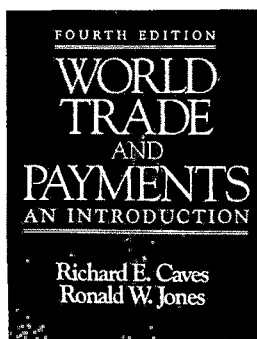
**US:** Maxwell House, Fairview Park, Elmsford, New York 10523 (914) 592-7700

**UK:** Headington Hill Hall, Oxford OX3 0BW, England

---

# Texts you can count on...

---



*New edition!*

## **WORLD TRADE AND PAYMENTS**

### **An Introduction**

*Fourth Edition*

**Richard E. Caves and Ronald W. Jones**

Thoroughly updated, this revision of the leading text for international economics features a reworked, up-to-date treatment of trade theory and strengthened coverage of international finance. And this clear, modern account is now more useful than ever — with supportive chapter-ending exercises and problems new to this edition.

#132276/cloth/544 pages

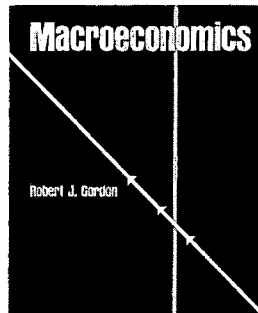
*New!*

## **MANAGERIAL ECONOMICS**

**Robert F. Rooney**

This comprehensive new text is more than merely a price theory text disguised as a managerial economics text. Rooney's fresh approach presents microeconomics in a more useful way for business students. The text features worked-out cases that show students how theory can be applied to practical business decisions and problems. Calculus is used where appropriate, and numerous problems are included at the end of each chapter.

#755966/cloth/500 pages/with Instructor's Manual



*Highly successful!*

## **MACROECONOMICS**

*Third Edition*

**Robert J. Gordon**

Users across the nation agree that Gordon's Third Edition is the best text — with the best package — for the intermediate macro course. Simplified presentation of inflation and unemployment as well as the latest data available enhance this widely-used text's superb integration of modern theory with numerous real-world cases.

#321079/cloth/720 pages/1984/with Instructor's Manual, Student Workbook, and *The Gordon Update* newsletter

## **MICROECONOMIC THEORY AND APPLICATIONS**

Edgar K. Browning and  
Jacqueline M. Browning

#112232/1983/cloth  
604 pages/with Instructor's  
Manual and Study Guide

## **PUBLIC SECTOR ECONOMICS**

*Second Edition*

Robin W. Boadway and  
David E. Wildasin

#100528/1984/cloth  
480 pages

## **ENVIRONMENTAL ECONOMICS AND POLICY**

Paul B. Downing  
#191809/1984/cloth  
350 pages

---

# from Little, Brown.

College Division • 34 Beacon Street • Boston, Massachusetts 02106

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

*The eighth—and election year—edition  
of the highly successful reader!*

**“An excellent book of supplemental readings  
for an introductory macroeconomics course.  
The articles are more current and politically  
diverse than most readings in the field.”  
—AAAS Science Books & Films**

# **INTRODUCTORY MACROECONOMICS 1984–85**

## **Readings on Contemporary Issues**

Edited by PETER D. McCLELLAND

The eighth in a series of highly successful and widely adopted annual readers for all introductory economics courses, *Introductory Macroeconomics 1984–85* contains more than fifty articles by leading economists published in periodicals as recently as May 1984. The articles are chosen to reflect the entire spectrum of economic opinion, from left-wing to right-wing, from liberal to conservative, from Keynesian to monetarist to supply-side advocate. Writers featured in this past year's reader included well-known commentators on macroeconomic problems such as Kenneth Arrow, Barbara R. Bergmann, Karl Brunner, Milton Friedman, Henry A. Kissinger, Joseph A. Pechman, Paul Craig Roberts, Paul Samuelson, David A. Stockman, and Paul A. Volcker.

The provocative articles and essays are selected from a diverse range of popular and scholarly sources, including *The Wall Street Journal*, *The Economist*, *Time*, *The Progressive*, *Dollars & Sense*, *New Republic*, *Economic Outlook USA*, *The New York Times*, *Challenge*, *U.S. News & World Report*, *Society*, *Newsweek*, Federal Reserve publications, and other major magazines and journals. The issues covered in *Introductory Macroeconomics 1984–85* include: inflation and unemployment, monetary and fiscal policy, institutional changes in money and banking, the growing federal deficit and proposals to reduce it, tax-based incomes policy, economic growth and reindustrialization, and poverty, welfare, and income distribution. **\$9.95 softcover**

— — — — — Send for a free examination copy today — — — — —

CORNELL UNIVERSITY PRESS  
P.O. Box 250, Ithaca, New York 14851  
Please send one copy of *Introductory  
Macroeconomics 1984–85* edited by  
Peter D. McClelland

I will be considering it for adoption in  
my course \_\_\_\_\_  
which has an expected enrollment of  
\_\_\_\_\_. The text I am now using is  
\_\_\_\_\_.

Name \_\_\_\_\_

Institution \_\_\_\_\_

Department \_\_\_\_\_

Street \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_ Zip \_\_\_\_\_

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

xxx

Look to St. Martin's Press  
for the latest word  
on major issues  
in contemporary economics

**Free Market Economics**  
**A CRITICAL APPRAISAL**

ANDREW SCHOTTER, New York University

A careful analysis of the free-market economics argument, this book defines the basic assumptions of the argument and identifies their roots in the history of economic thought. Schotter reviews the major criticisms of the argument and raises some serious doubts about the blind advocacy of free market solutions to current social problems.

Paperbound. 147 pages.  
Just Published

**The Politics of International  
Economic Relations**  
**Third Edition**

JOAN EDELMAN SPERO, Senior Vice  
President, American Express Company

A clearly written, carefully focused examination of the political dynamics of international economics, this book addresses all the major topics of concern in the field, including money, trade, multinational corporations, North-South relations, economic aid, and oil. This Third Edition provides new or expanded coverage of such important topics as protectionism, the declining power of OPEC, the consequences of Third World debt, and the impact of high U.S. interest rates and the overvalued dollar. A glossary of key economic terms is now included at the back of the book.

Paperbound. 400 pages (probable).  
February 1985

**The Contemporary  
International Economy**  
**A READER**

**Second Edition**

JOHN ADAMS, University of Maryland,  
College Park

Intended as a supplement for courses in international economics and business, this book brings together 33 readings on the major issues in international economics today. The book provides a broad range of viewpoints, with selections by such noted economists as Martin Feldstein, Robert Baldwin, Herbert Grubel, Mordechai Kreinin, and Peter Drucker. The book is fully up to date: all but four of the readings are new to this edition.

Paperbound. 550 pages (probable).  
February 1985

To request a complimentary examination copy of any of these titles, please write us on your college letterhead specifying your course title, present text, and approximate enrollment. Address your request to:

**ST. MARTIN'S**  
**PRESS** Department JR  
175 Fifth Avenue  
New York, N.Y. 10010

# AMERICAN ECONOMIC ASSOCIATION

## 1985 ANNUAL MEMBERSHIP RATES

### Membership includes:

—a subscription to both *The American Economic Review* (quarterly) plus *Papers and Proceedings* and the *Journal of Economic Literature* (quarterly).

- Regular members with annual incomes of \$30,000 or less ..... \$35.00
- Regular members with annual incomes above \$30,000 but no more than \$40,000 ..... \$42.00
- Regular members with annual incomes above \$40,000 ..... \$49.00
- Junior members (available to registered students for three years only).

Student status must be certified by your major professor or school registrar ..... \$17.50

- In Countries other than the U.S.A., Add \$11.00 to cover postage.
- Family members (persons living at the same address as a regular member, additional memberships without subscription to the publications of the Association) ..... \$7.00

Please begin my issues with:

☐ March

☐ June  
(Includes Papers  
and Proceedings)

☐ September

☐ December

First Name and Initial	Last Name	Suffix
Address Line 1		
Address Line 2		
City		
State or Country	Zip/Postal Code	

**MAJOR FIELDS (TWO ONLY)**  
LIST FIELDS WITH WHICH  
YOU CURRENTLY IDENTIFY.  
SELECT FIELD CODE FROM JEL,  
"Classification System  
for Books."

PLEASE TYPE OR PRINT INFORMATION ABOVE; PLEASE SEND CHECK OR MONEY ORDER PAYABLE IN U.S. DOLLARS. CANADIAN AND FOREIGN PAYMENTS MUST BE IN THE FORM OF A U.S. DOLLAR DRAFT ON A NEW YORK BANK.

Endorsed by (AEA member) \_\_\_\_\_

**Below for Junior Members Only**

I certify that the person named above is enrolled as a student at \_\_\_\_\_

\_\_\_\_\_  
Authorized Signature

PLEASE SEND WITH PAYMENT TO:

**AMERICAN ECONOMIC ASSOCIATION**  
1313 21ST AVENUE SOUTH, SUITE 809  
NASHVILLE, TENNESSEE 37212-2786  
U.S.A.

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*



# JOB OPENINGS FOR ECONOMISTS

Available only to AEA members and institutions that agree to list their openings.

## Annual Subscription Rates

U.S.A., Canada, and Mexico (first class): \$15.00, regular AEA members and institutions  
\$ 7.50, junior members of AEA  
All other countries (air mail): \$22.50, regular AEA members and institutions  
\$15.00, junior members of AEA

Please begin my issues with:

☐ February ☐ April ☐ June ☐ August ☐ October ☐ December

Name \_\_\_\_\_  
First Middle Last

Address \_\_\_\_\_

\_\_\_\_\_  
City State/Country Zip/Postal Code

Check one:

- ☐ I am a member of the American Economic Association.  
☐ I would like to become a member. My application and payment are enclosed.  
☐ (For institutions) We agree to list our vacancies in JOE.

Send payment (U.S. currency only) to:

THE AMERICAN ECONOMIC ASSOCIATION  
1313 21st Avenue South  
Nashville, Tennessee 37212

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

# Oxford

## Order and Conflict in Contemporary Capitalism

**Studies in the Political Economy of Western European Nations**

Edited by JOHN H. GOLDTHORPE, *Nuffield College, Oxford*. This comparative anthology examines the methods used by Western countries to manage political conflict over the distribution of economic and social resources since 1945. Bringing together major studies by leading economists, political scientists, sociologists, and historians, the collection offers a new basis for assessing the relative merits and weaknesses of corporatist and other settlement strategies.

1985 352 pp. cloth \$34.95 paper \$15.95

## The Way People Work

**Job Satisfaction and Customer Service**

CHRISTINE HOWARTH. What makes a job satisfying? Does greater job satisfaction lead to increased efficiency and, if so, is the extra cost worthwhile? How can managers and employees improve the quality of working life? Using detailed case studies of various American and European organizations, this practical guide explains how workplaces can be personally—and economically—rewarding for managers and employees alike.

1984 224 pp. \$23.95

## The Co-operative Game Theory of the Firm

MASAHIKO AOKI, *Institute of Economics, Kyoto University, Japan, and Stanford University*. This highly original work challenges the widely-held view of the firm as a mysterious "black box," operating solely to maximize profit for the shareholders. Instead, Aoki proposes an entirely new economic model in which the ideal firm is a coalition of shareholders and employees, with its market behavior and internal distribution the result of co-operative bargaining.

1985 288 pp.; 9 figs. \$29.95

## Putting People First

**Sociological Dimensions of Rural Development**

Edited by MICHAEL M. CERNEA. This useful anthology addresses the social and practical considerations involved in the design and implementation of effective rural development projects. The contributors, studying topics ranging from the social organization of the productive system to the role of the World Bank, provide solid advice and recommendations for incorporating regional sociological factors into the design of any successful development project. (*A World Bank Publication*)

1985 400 pp. \$24.95

## Winners and Losers in Colombia's Economic Growth of the 1970s

MIGUEL URRUTIA, *United Nations University, Tokyo*. Most experts believe that Colombia's economic growth during the 1970s did not improve the distribution of income. Here, Urrutia draws on a wide array of statistical data to reach a different conclusion: that income distribution during this period did not, in fact, deteriorate and that the real incomes of the poorest workers actually improved markedly. (*A World Bank Publication*)

1985 160 pp. \$19.95

## Input-Output Economics

*Second Edition*

WASSILY LEONTIEF, *Institute for Economic Analysis, New York University*. In 1966, Wassily Leontief was awarded the Nobel Prize for his model of Input-Output economics. This collection of papers provides the only comprehensive introduction to the model he has written, and is now thoroughly revised—12 of the 18 papers are new to this edition. The book begins with non-technical articles on the theory of Input-Output economics and progresses to more technical discussions and then to specific applications of the theory.

May 1985 320 pp.; illus. cloth \$29.95 paper \$14.95

## Leading Issues in Economic Development

*Fourth Edition*

GERALD M. MEIER, *Stanford University*. Completely revised and updated to reflect the current mood of stock taking and reassessment, this unique book brings order to the diffuse subject of economic development while maintaining a variety of viewpoints and different perspectives. The author uses a broad range of theoretical, applied, and policy materials in emphasizing the strategic policy issues which accelerate the development of Third World nations.

1984 768 pp.; 45 illus. paper \$19.95

## The World Crisis in Education

*The View From the Eighties*

PHILIP H. COOMBS, *International Council for Educational Development*. In this sequel to his highly acclaimed book, *The World Educational Crisis*, Philip Coombs identifies significant educational trends that have developed over the past 20 years and examines critical future issues destined to confront all nations, with varying intensity, over the next 15 years. Possible ways of coping with these issues are explored through educational reforms, innovations, and new forms of international cooperation.

1985 384 pp.; 25 illus. cloth \$24.95 paper \$10.95

## Taxation for Development

*Principles and Applications*

STEPHEN R. LEWIS, *Williams College*. This introduction to the analysis of tax policy in developing countries focuses on open economies and the ways in which openness to international trade and to movements of capital and skilled labor influence the scope for, and the effects of, tax policy. More than 50 policy problems—based on actual case studies from a variety of countries—allow the reader to apply principles developed in theory to actual tax policy issues.

1985 256 pp.; illus. cloth \$29.95 paper \$12.95

*Prices and publication dates are subject to change.*

To order send check or money order to Assistant Marketing Manager,  
Humanities and Social Sciences

**Oxford University Press**

200 Madison Avenue, New York, NY 10016

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

# The Economics Institute

Boulder, Colorado

---

- PREPARATION FOR MASTERS AND DOCTORAL DEGREE PROGRAMS IN ECONOMICS, BUSINESS, AND ADMINISTRATION.
  - POSTGRADUATE DIPLOMA PROGRAMS IN RELATED SPECIALIZATIONS.
  - AN ESTABLISHED REPUTATION FOR ACADEMIC EXCELLENCE.
- 

25 years of specialized service in  
international education.

---

Sponsored by the American  
Economic Association



For further information write:

The Director  
Economics Institute  
Campus Box 259  
University of Colorado  
Boulder, Colorado 80309  
(303) 492-7337  
Telex: 45-0385

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

## Econometric/Statistical Package

# xSTAT™

### Mainframe Capability, Accuracy and Speed for Your IBM Personal Computer

---

**xSTAT:****A Reliable, Economical Alternative to Mainframe Computing**

---

- Is designed for the analysis of **VERY LARGE** as well as small data sets. It has no limit on sample size (total number of observations); and allows up to 50 variables per regression.
- Can generate and save an **X'X product matrix** the first time a data set is used. Subsequent regressions can be made from it to eliminate repetitive readings of input data normally required in statistical packages—very time-consuming when disk inputs are involved.
- Supports Intel 8087 80-bit Math Co-processor for **FAST NUMBER CRUNCHINGS** (has software emulation for systems without the co-processor to attain same accuracy). Double precision is used in matrix operations.
- Produces estimated coefficients that are identical to mainframe TSP results to four significant digits.
- Executes at **HIGH SPEED** when an **X'X matrix** is used to estimate regressions even for very large samples—it takes only seconds to complete a run.
- The output can be directed to disk, printer or screen. The display can be 132 or 80 columns wide and the statistics and coefficients can be printed in F-format (e.g., 123.456) or E-format (1.23456 + E02).
- Works in either batch or prompt mode; requires no special command language. **EASY TO USE.**

#### THE PACKAGE INCLUDES:

##### Data Reformatting and Transformation Program

- Converts formatted data into unformatted data and vice versa; aggregates over records; merges files; accepts data downloaded from mainframe or generated by other programs including Lotus 1-2-3, Visicalc, dBase II, WordStar, other word processors and editors.
- Allows selection of subsamples based on value range or sequential order. Converts fractions to decimal numbers (e.g., stock quotes,  $32\frac{1}{8} = > 32.125$ ).
- Has 14 mathematical/logical transformation functions for vector multiplication, division, addition, subtraction, logarithm, exponent, lagged variables, first difference, dummy variables, and time trend.

##### Descriptive Statistics

- Minimum, maximum, mean, standard deviation, sum of squares, correlation coefficients, **CROSSTABS**, etc.

##### Ordinary Least Squares (Multiple regressions) and Two-stage Least Squares

- Generates variance/covariance matrix of coefficients, standard error, t-statistic, R-square, adjusted R-square, Durbin-Watson statistic and F-statistic. Calculates predicted values, residuals, and produces residual plot.
- Allows subsample selection; can also estimate regressions from product moment matrix (contingency table).

##### Pooling Cross-Section and Time-Series Data

- Estimates regressions for samples with multi-period observations (e.g., panel study). Automatic stacking of data matrix (diagonal block matrix). Estimates coefficients and F-statistics for testing various hypotheses of (1) a common intercept and slope; (2) a common slope and different intercepts; (3) a common intercept and different slopes; (4) different intercepts and different slopes.

#### SYSTEM REQUIREMENTS

**IBM PERSONAL COMPUTER** (PC or XT), COMPAQ or IBM PC compatible systems: 1 floppy disk drive or a hard disk system; DOS 1.1 or 2.0 and 192K memory.

**PRICE:** \$250 for the complete package. (Accept checks, Visa & MasterCard. Or order through your local dealer.)

#### RELATED PRODUCT

Spreadsheet Forecasting Templates for Lotus 1-2-3: reads the estimated coefficients and inverse **X'X** matrix, saved optionally by xSTAT, directly into 1-2-3 for making "what if" forecasts and producing graphs. No keyboard data entry is needed. The templates are imbedded with formulae for estimating the standard error and confidence interval for the forecasted dependent variable that changes as the forecasted independent variables are varied. **PRICE \$100**

---

### MING TELECOMPUTING INC.

Telecommunications and Statistics for Microcomputers

P.O. BOX 101

Lincoln Center, MA 01773

(617) 259-0391

(SAMPLE OUTPUTS ARE AVAILABLE UPON REQUEST.)

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

# CSWEP

*The Committee on the Status of Women in the Economics Profession*

**CSWEP**, an arm of the American Economic Association, publishes a thrice-yearly newsletter to let you know all the news - good or bad - about women in the economics profession.

The newsletter carries items about:

*research on sex role issues*

*new publications*

*hiring (or non-hiring) of women economists*

*conferences for or about women*

*what other professional women's groups are doing to  
further women's interests in their disciplines*

*developments in government and industry*

**CSWEP** speaks up on behalf of women economists in hiring, research and governmental policies.

**CSWEP** represents women's point of view in the committee work of the American Economic Association. It makes an annual report to AEA on the status of women economists.

**CSWEP** is a presence at annual meetings of the AEA and of the regional economics associations. It sponsors sessions at these meetings, where research by and about women can get an audience.

**CSWEP** publishes a **ROSTER OF WOMEN ECONOMISTS** for purposes of communication and job placement. Dues-paying members receive the latest roster, which lists women economists by and where they teach or work, by their specialty in economics, and by city, as well as alphabetically.

**CSWEP** is the voice of women economists when coalitions of professional women join to advance sex equality in professional life.

To become a dues-paying member of **CSWEP**  
send this with a check for \$15 (tax deductible) made out to **CSWEP** to:  
**CSWEP**, 3065 Fermanagh Dr., Tallahassee, FL 32308

NAME \_\_\_\_\_

MAILING ADDRESS \_\_\_\_\_

CITY, STATE, ZIP \_\_\_\_\_

Check here if currently an AEA member ☐

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

# A New Standard Of Excellence



**MACROECONOMICS: The Dynamics Of Theory And Policy—  
an exciting and successful new text by William J. Boyes:**

- Provides a clear integration of macro theory and policies within the historical context of their development
- Progresses rapidly to a development of the Aggregate Demand/Aggregate Supply framework and then uses it to unify the treatment of theory
- Devotes an entire chapter to an examination of alternative expectations schemes—including rational and adaptive expectations models
- Applies the unifying AD/AS Model to examine such current issues as tax policy, budgets and deficits, Social Security, declining productivity growth, the political business cycle, and the Third World Debt
- Presents a balanced treatment of current controversies—including rules versus discretion, the explanation of stagflation, contract- and information-based theories of the business cycle, and supply-side economics

**For more information on this text and other South-Western titles,  
contact:**

**Lew Gossage  
COLLEGE DIVISION  
SOUTH-WESTERN PUBLISHING CO.  
5101 Madison Road  
Cincinnati, OH 45227.**

**MATHEMATICAL AND STATISTICAL PROGRAMMING PACKAGE FOR YOUR IBM PC**  
FAST • EASY TO USE • POWERFUL

# GAUSS™

**YOU'VE NEVER SEEN ANYTHING LIKE IT!**

**GAUSS** is a sophisticated mathematical and statistical programming package for the IBM PC and compatibles. It combines speed, power, and ease of use in one amazing program.

**GAUSS** allows you to do essentially anything you can do with a mainframe statistical package — and a lot more.

Personal computers are friendly, convenient, and inexpensive. So is **GAUSS**. **GAUSS** is not just a stripped-down mainframe program. **GAUSS** has been designed from the ground up to take advantage of all of the conveniences of a personal computer. After trying **GAUSS**, you may never use a mainframe again.

**GAUSS** comes with programs written in its matrix programming language that allow you to do most econometric procedures, including OLS, 2SLS, 3SLS, PROBIT, LOGIT, MAXIMUM LIKELIHOOD, and NON-LINEAR LEAST SQUARES.

In the current version, **GAUSS** will accept up to 90 variables in a regression. There is no limit on the number of observations.

**GAUSS** will do a regression with 10 independent variables and 800 observations in under 4 seconds — and with 50 variables and 10,000 observations in under 18 minutes. It will compute the maximum likelihood estimates of a binary logit model, with 10 variables and 1,000 observations, in 1-2 minutes, depending upon the number of iterations required.

**GAUSS** allows you to do complicated statistical procedures that you would never imagine trying on a mainframe. It is easy to program almost any routine, and **GAUSS** is so fast that it can do almost any job. But the nicest thing of all is that the cost of time on your personal computer is essentially zero!

**GAUSS** is an excellent teaching tool. It makes programming easy and allows students to focus on concepts and techniques.

If you can write it mathematically, you can write it in **GAUSS**. Furthermore, you can write it in **GAUSS** almost exactly the way you would write it mathematically.

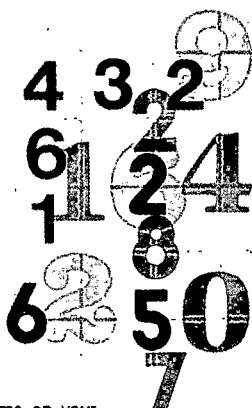
**GAUSS** is 10-15 times faster than other programs that use the 8087, and 15-100 times faster than other programs that do not use the 8087.

As in APL, single statements in **GAUSS** can accomplish what might take dozens of lines in another language. However, **GAUSS** provides you with additional powerful numerical operators and functions — especially for statistics and the solution of linear equations — that are not found in APL. And, of course, the syntax in **GAUSS** is much more natural (for most of us) than that in APL.

**GAUSS** has state-of-the-art numerical routines and random number generators.

**GAUSS** is extremely accurate. It allows you to do an entire regression in 19 digit accuracy. It will compute the Longley benchmark coefficients in 5 hundredths of a second with an average of 11 correct digits! (Try that on a mainframe!)

**GAUSS**, with its built-in random number generators and powerful functions and operators, is an excellent tool for doing simulations.



## **GAUSS and the 8087 NUMERIC DATA PROCESSOR GIVE YOU MINICOMPUTER PERFORMANCE ON YOUR DESKTOP.**

### **SPECIAL INTRODUCTORY OFFER**

With 30 Day Money

Back Guarantee ..... Reg. 395.00 **\$250.00**

**GAUSS** requires an IBM PC with at least 256K RAM, an 8087 NDP, 1 DS/DD disk drive, DOS 2.0 (or above).

IBM is trademark of IBM Corporation.

Call or Write

**APPLIED  
TECHNICAL  
SYSTEMS**

P.O. Box 6487, Kent, WA 98064  
(206) 631-6679

**Come and get a demonstration of GAUSS at our booth in the exhibitor's area at the AEA meetings in Dallas, December 28-30, 1984.**



# INDEX OF ECONOMIC ARTICLES

prepared under the auspices of  
*The Journal of Economic Literature*  
of the  
*American Economic Association*

- ✓ Each volume in the **Index** lists articles in major economic journals and in collective volumes published during a specific year.
- ✓ Most of the **Index's** volumes also include articles of testimony from selected congressional hearings in government documents published during the year.
- ✓ No other single reference source covers as many articles classified in economic categories as the **Index**.
- ✓ The 1977 volume contains over 10,500 entries.

## Currently available are:

Volume	Year Covered
XI	1969
XII	1970
XIII	1971
XIV	1972
XV	1973
XVI	1974
XVII	1975
XVIII	1976
XIX	1977
XX	1978
XXI	1979



*an  
indispensable  
tool for...*

**ECONOMISTS  
REFERENCE LIBRARIANS  
RESEARCHERS  
TEACHERS  
STUDENTS  
AUTHORS**

Future volumes will be published regularly  
to keep the series as current as possible.

**Price:** \$50.00 per volume (special 30% discount to  
AEA members)

Distributed by

**RICHARD D. IRWIN, INC.** Homewood, Illinois  
60430

# ECONOMICS

**1985**

**NEW:**

- Rosen  
**PUBLIC FINANCE**
- Blair & Kaserman  
**ANTITRUST  
ECONOMICS**

**REVISIONS:**

- Maurice & Smithson  
**MANAGERIAL  
ECONOMICS**, 2nd Edition

- Reynolds  
**MACROECONOMICS:  
Analysis and Policy**  
5th Edition  
**MICROECONOMICS:  
Analysis and Policy**  
5th Edition

- Tullock & McKenzie  
**THE NEW WORLD  
OF ECONOMICS:  
Explorations into  
the Human  
Experience**, 4th Edition

- Shepherd  
**PUBLIC POLICIES  
TOWARD BUSINESS**  
7th Edition

- Begin & Beal  
**THE PRACTICE  
OF COLLECTIVE  
BARGAINING**  
7th Edition

- Rowan  
**READINGS IN  
LABOR ECONOMICS  
AND LABOR  
RELATIONS**  
5th Edition

- Bornstein  
**COMPARATIVE  
ECONOMIC SYSTEMS:  
Models and Cases**  
5th Edition

**1984**

- Wonnacott  
**MACROECONOMICS**  
3rd Edition

- Seo  
**MANAGERIAL  
ECONOMICS: Text,  
Problems, and  
Short Cases**  
6th Edition

Examination copies for adoption consideration  
available upon request. Please indicate  
course title and text presently used.



**Richard D. Irwin, Inc.**  
Homewood, IL 60430



# The American Economic Review

## PAPERS AND PROCEEDINGS

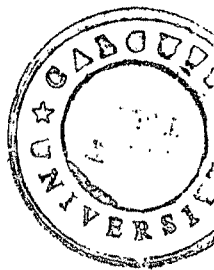
OF THE

Ninety-Seventh Annual Meeting

OF THE

AMERICAN ECONOMIC ASSOCIATION

Dallas, Texas, December 28-30, 1984



Program Arranged by Charles P. Kindleberger

Papers and Proceedings Edited by John G. Riley and Wilma St. John

MAY 1985

# THE AMERICAN ECONOMIC ASSOCIATION

●Published at George Banta Co., Inc., Menasha, Wisconsin. The publication number is ISSN 0002-8282.

●*THE AMERICAN ECONOMIC REVIEW* including four quarterly numbers, the *Proceedings* of the annual meetings, the *Directory*, and *Supplements*, is published by the American Economic Association and is sent to all members five times a year: March; May; June; September; December.

Regular member dues for 1985, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature* are as follows:

- \$35.00 if annual income is \$30,000 or less;
- \$42.00 if annual income is above \$30,000, but no more than \$40,000;
- \$49.00 if annual income is above \$40,000.

Nonmember subscriptions will be accepted only for both journals: Institutions (libraries, firms, etc.), \$100 a year; individuals, \$65.00. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.

In countries other than the United States, add \$11.00 to the annual rates above to cover extra postage.

●Correspondence relating to the *Directory*, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

●Second-class postage paid at Nashville, Tennessee and at additional mailing offices, Printed in U.S.A.

●Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786.

Founded in 1885

## Officers

### *President*

CHARLES L. SCHULTZE

The Brookings Institution

### *President-Elect*

CHARLES P. KINDLEBERGER

Massachusetts Institute of Technology

### *Vice Presidents*

ZVI GRILICHES

Harvard University

ROBERT L. HEILBRONER

New School for Social Research

### *Secretary*

C. ELTON HINSHAW

Vanderbilt University

### *Treasurer*

RENDIGS FELS

Vanderbilt University

### *Managing Editor of The American Economic Review*

ROBERT W. CLOWER

University of California-Los Angeles

### *Managing Editor of The Journal of Economic Literature*

MOSES ABRAMOVITZ

Stanford University

## Executive Committee

### *Elected Members of the Executive Committee*

ANN F. FRIEDLAENDER

Massachusetts Institute of Technology

JOSEPH E. STIGLITZ

Princeton University

WILLIAM D. NORDHAUS

Yale University

A. MICHAEL SPENCE

Harvard University

VICTOR R. FUCHS

Stanford University

JANET L. NORWOOD

Bureau of Labor Statistics

### *EX OFFICIO Members*

GARDNER ACKLEY

The University of Michigan

W. ARTHUR LEWIS

Princeton University

# THE AMERICAN ECONOMIC REVIEW

---

VOL. 75 NO. 2

MAY 1985

---

*PAPERS AND PROCEEDINGS*

OF THE

*Ninety-Seventh Annual Meeting*

OF THE

AMERICAN ECONOMIC ASSOCIATION

Dallas, Texas

December 28-30, 1984

*Program Arranged by* Charles P. Kindleberger

*Papers and Proceedings Edited by* John G. Riley and Wilma St. John

Copyright © AMERICAN ECONOMIC ASSOCIATION, 1985



## CONTENTS

Editors' Introduction .....	<i>John G. Riley and Wilma St. John</i>	vii
-----------------------------	---	-----

## PAPERS

<b>Richard T. Ely Lecture</b>		
Economics in Theory and Practice .....	<i>Sir Alec Cairncross</i>	1
<b>Issues in the Economics of <i>R&amp;D</i></b>		
Post-Entry Competition in the Plain Paper Copier Market .....	<i>Timothy F. Bresnahan</i>	15
<i>R&amp;D</i> Appropriability, Opportunity Market Structure: New Evidence on Some Schumpeterian Hypotheses .....	<i>Richard C. Levin, Wesley M. Cohen, and David C. Mowery</i>	20
Patent Licensing and <i>R&amp;D</i> Rivalry .....	<i>Carl Shapiro</i>	25
<b>Labor Contracts and Macroeconomic Performance</b>		
Nominal Wage-Price Rigidity as a Rational Expectations Equilibrium .....	<i>Costas Azariadis and Russell Cooper</i>	31
Wage Flexibility in the United States: Lessons from the Past .....	<i>Daniel J. B. Mitchell</i>	36
Profit Sharing as Macroeconomic Policy .....	<i>Martin L. Weitzman</i>	41
<b>The Consequences of Deregulation in the Transportation and Telecommunications Sectors</b>		
The Regulatory Transition .....	<i>John R. Meyer and William B. Tye</i>	46
"Let Them Make Toll Calls": A State Regulator's Lament .....	<i>Roger G. Noll</i>	52
Intercity Transportation Route Structures under Deregulation: Some Assessments Motivated by the Airline Experience .....	<i>Steven A. Morrison and Clifford Winston</i>	57
<b>The End of the Great Boom and the Breakdown of Bretton Woods: Was it a Coincidence?</b>		
Macroeconomic Stability and Flexible Exchange Rates .....	<i>John F. O. Bilson</i>	62
Reflections on the Exchange Rate System .....	<i>J. Carter Murphy</i>	68
On the System in Bretton Woods .....	<i>John Williamson</i>	74
<b>Economic Education: The Use of Computers</b>		
Computer Applications in Pre-College Economics .....	<i>John M. Sumansky</i>	80
Macro Simulations for PCs in the Classroom .....	<i>Karl E. Case and Ray C. Fair</i>	85
Cost Effectiveness of Computer-Assisted Economics Instruction .....	<i>Darrell R. Lewis, Bruce R. Dalgaard, and Carol M. Boyer</i>	91
<b>The Deregulation of Banking in the United States</b>		
Legislative Construction of the Monetary Control Act of 1980 .....	<i>Richard H. Timberlake, Jr.</i>	97
Deregulation and Monetary Reform .....	<i>Leland B. Yeager</i>	103
Speculation, Deregulation, and the Interest Rate .....	<i>Leonard A. Rapping and Lawrence B. Pulley</i>	108
<b>Yesteryears' Long-Range Projections: A Retrospective</b>		
1985 Projections of the New York Metropolitan Region Study .....	<i>Dick Netzer</i>	114
Long-Term Forecasts in International Economics .....	<i>William R. Cline</i>	120
The Economic Thought of George Orwell .....	<i>Jennifer Roback</i>	127

**The After-Keynes Cambridge Contributions to Theory**

- Hamlet without the Prince: Cambridge Macroeconomics without Money . . . . . *J. A. Kregel* 133  
 Cambridge Price Theory: Special Model or General Theory of Value? . . . . . *Bertram Schefold* 140  
 Joan Robinson's Critique of Equilibrium: An Appraisal . . . . . *E. Roy Weintraub* 146

**Open and Sealed-Bid Auctions**

- Auction Theory with Private Values . . . . . *Eric S. Maskin and John G. Riley* 150  
 Empirical Testing of Auction Theory . . . . . *Robert G. Hansen* 156  
 Experimental Development of Sealed-Bid Auction Theory: Calibrating Controls for Risk  
 Aversion . . . . . *James C. Cox, Vernon L. Smith, and James M. Walker* 160

**Frontiers in Demographic Economics**

- Modes of Thought in Economics and Biology . . . . . *Paul A. Samuelson* 166  
 The New Economics of Labor Migration . . . . . *Oded Stark and David E. Bloom* 173  
 New Evidence on the Timing and Spacing of Births . . . . .  
 . . . . . *James J. Heckman, V. Joseph Holtz, and James R. Walker* 179

**Perspective on the External Debt Situation**

- International Debt: From Crisis to Recovery? . . . . . *William R. Cline* 185  
 Latin American Debt: Lessons and Pending Issues . . . . . *Eduardo Wiesner* 191

**The Use and Abuse of Econometrics**

- Data and Econometricians—The Uneasy Alliance . . . . . *Zvi Griliches* 196  
 The Loss Function has been Misplaced: The Rhetoric of Significance Tests . . . . .  
 . . . . . *Donald N. McCloskey* 201  
 Macroeconomic Modeling and the Theory of the Representative Agent . . . . . *John Geweke* 206

**Modeling Intercountry Linkages**

- An Integrated Accounting Matrix for Canada and the United States . . . . .  
 . . . . . *Jacob Cohen and Steven Husted* 211  
 Modeling United States—Mexico Economic Linkages . . . . .  
 . . . . . *Clark W. Reynolds and Robert McCleery* 217  
 New Developments in Project LINK . . . . . *L. R. Klein* 223

**In Honor of Stephen H. Hymer: The First Quarter Century of the Theory of Foreign Direct Investment**

- The Influence of Hymer's Dissertation on the Theory of Foreign Direct Investment . . . . .  
 . . . . . *John H. Dunning and Alan M. Rugman* 228  
 Multinational Enterprise, Internal Governance, and Industrial Organization . . . . *David J. Teece* 233  
 Stephen Hymer and Public Policy in LDCs . . . . . *Donald J. Lecraw* 239

**Human Capital and Culture: Analyses of Variations in Labor Market Performance**

- Religion and the Earnings Function . . . . . *Nigel Tones* 245  
 Cultural Differences in Labor Force Participation among Married Women . . . . .  
 . . . . . *Cordelia W. Reimers* 251  
 Peddlers Forever?: Culture, Competition, and Discrimination . . . . .  
 . . . . . *William A. Darity, Jr. and Rhonda M. Williams* 256

**Is Gender Equality Advancing in the Workplace?**

- Women Production Workers: Low Pay and Hazardous Work . . . . . *Janis Barry* 262  
 Executive Compensation: Female Executives and Networking . . . . .  
 . . . . . *Robin L. Bartlett and Timothy I. Miller* 266  
 Longitudinal Changes in Salary at a Large Public University: What Response to Equal Pay  
 Legislation? . . . . . *Sharon Bernstein Megdal and Michael R. Ransom* 271  
 Sex Role Socialization and Labor Market Outcomes . . . . *Mary E. Corcoran and Paul N. Courant* 275

**The Pacific Challenge for World Economic Leadership**

- Pacific Protagonist: Implications of the Rising Role of the Pacific . . . *Staffan Burenstam Linder* 279  
 Is There Need for Economic Leadership: Japanese or U.S.? . . . *W. W. Rostow* 285

**The Theory of Economic Organizations**

- Human Fallibility and Economic Organization . . . . . *Raaj Kumar Sah and Joseph E. Stiglitz* 292  
 Learning from Experience in Organizations . . . . .  
 . . . . . *Scott R. Herriott, Daniel Levinthal, and James G. March* 298  
 Informational Structure of the Firm . . . . . *Kenneth J. Arrow* 303

**Industrial Policy in France**

- State and Industry in France, 1750–1914 . . . . . *Caglar Keyder* 308  
 French Industrial Policy under the Socialist Government . . . . . *Bela Balassa* 315

**Economic History: A Necessary though not Sufficient Condition for an Economist**

- Maine and Texas . . . . . *Kenneth J. Arrow* 320  
 Is History Stranger than Theory?: The Origins of Telephone Separations . . . . .  
 . . . . . *Peter Temin and Geoffrey Peters* 324  
 Economic History and Economics . . . . . *Robert M. Solow* 328  
 CLIO and the Economics of QWERTY . . . . . *Paul A. David* 332

**Credit and Economic Instability**

- Portfolio Choice and the Debt-to-Income Relationship . . . . . *Benjamin M. Friedman* 338  
 Stability and Instability in the Debt-Income Relationship . . . . . *Robert Pollin* 344  
 Private Credit Demand, Supply, and Crunches: How Different are the 1980's? . . . . .  
 . . . . . *Albert M. Wojnilower* 351

**Macroeconomic Analysis of Leading Interwar Authorities**

- Marriner S. Eccles, Chairman of the Federal Reserve Board . . . . . *L. Dwight Israelsen* 357  
 Rudolf Hilferding: The Dominion of Capitalism and the Dominion of Gold . . . . .  
 . . . . . *William A. Darity, Jr. and Bobbie L. Horn* 363  
 Korekiyo Takahashi and Japan's Recovery from the Great Depression . . . . .  
 . . . . . *Dick K. Nanto and Shinji Takagi* 369

**Risk Perception and Market Performance**

- Financial Risk and the Burdens of Contracts . . . *Herman B. Leonard and Richard J. Zeckhauser* 375  
 Are Individuals Bayesian Decision Makers? . . . . . *W. Kip Viscusi* 381  
 Ambiguity and Insurance Decisions . . . . . *Robin M. Hogarth and Howard Kunreuther* 386

**Uncertainty, Behavior, and Economic Theory**

- Origin of Predictable Behavior: Further Modeling and Applications . . . . . *Ronald A. Heiner* 391  
 Individual Rationality, Market Rationality, and Value Estimation . . . . .  
 . . . . . *Peter Knez, Vernon L. Smith, and Arlington W. Williams* 397  
 Knowledge, Uncertainty and Behavior . . . . .  
 . . . . . *Keith D. Wilde, Allen D. LeBaron, and L. Dwight Israelsen* 403

**International Finance**

- The Changing Environment of Central Bank Policy . . . . . *Alexandre Lamfalussy* 409



PROCEEDINGS

Minutes of the Annual Meeting .....	417
Minutes of the Executive Committee Meetings .....	418
Reports	
Secretary .....	C. Elton Hinshaw 423
Treasurer .....	Rendigs Fels 428
Finance Committee .....	Rendigs Fels 431
Managing Editor, <i>American Economic Review</i> .....	Robert W. Clower 432
Managing Editor, <i>Journal of Economic Literature</i> .....	Moses Abramovitz 438
Director, <i>Job Openings for Economists</i> .....	C. Elton Hinshaw 440
Census Advisory Committee .....	Richard E. Quandt 442
Representative to the National Bureau of Economic Research .....	Carl F. Christ 443
Policy and Advisory Board of the Economics Institute .....	Edwin S. Mills 445
Representative to UNESCO .....	Walter S. Salant 446
Committee on the Status of Women .....	Barbara R. Bergmann 448
Committee on U.S.-China Exchanges .....	Gregory C. Chow 454
Committee on Economic Education .....	W. Lee Hansen 455
<i>Ad Hoc</i> Committee on Financial Reporting and Changing Prices .....	Franco Modigliani 456

THE purpose of the American Economic Association, according to its charter, is the encouragement of economic research, the issue of publications on economic subjects, and the encouragement of perfect freedom of economic discussion. The Association as such takes no partisan attitude, nor does it commit its members to any position on practical economic questions. It is the organ of no party, sect, or institution. People of all shades of economic opinion are found among its members, and widely different issues are given a hearing in its annual meetings and through its publications. The Association, therefore, assumes no responsibility for the opinions expressed by those who participate in its meetings. Moreover, the papers presented are the personal opinions of the authors and do not commit the organizations or institutions with which they are associated.

## Editors' Introduction

This volume contains the *Papers and Proceedings* of the ninety-seventh annual meetings of the American Economic Association.

The *Proceedings* record the business activities of the Association in 1984; the annual membership meeting; the March and December meetings of the Association's officers and committees.

The *Papers* constitute the greater part of the volume. They comprise seventy-three contributions that fill roughly the same number of pages as two regular issues of the *American Economic Review*. About a year in advance, the Association's President-elect, acting as program chairman, decides on the topics for which sessions will be organized. This is done after consultation and comment, both volunteered and solicited, from a wide range of individuals. (A *Call for Papers* is published annually in the Notes section of the December issue of the *AER*.) The President-elect invites persons to organize these sessions. Each session organizer in turn invites several persons (usually two or three) to give papers on the theme of the session, and asks others to give comments on the papers. The program chairman decides at the time of organization which sessions are to be included in this volume. Space limitations restrict the number of printed sessions. This year we are printing 26 sessions, although a total of 107 sessions were sponsored, either solely by the American Economic Association or jointly with other allied societies. There has been no standard practice with regard to the publication of comments and discussions in the past. This year the President-elect chose to publish no comments, given the difficulty and the invidious task of choosing. He has arranged instead that the names and addresses of commentators be printed at the start of each session, permit-

ting readers especially interested in particular comments writing to the commentator for a copy of the discussion.

The guidelines under which papers are published in the *Papers and Proceedings* differ from those governing regular issues of the *Review*. First, the length of papers is strictly controlled. Except in unusual circumstances they must be no more than twelve typescript pages in three-paper sessions, and eighteen typescript pages in two-paper sessions. Second, papers are not subjected to any refereeing process. Third, their content and range of subject matter reflect the wishes of the President-elect to investigate and expose the current state of economic research and thinking. In most cases they are therefore exploratory and discursive, rather than formal presentations of original research.

While authors are encouraged to submit their manuscripts earlier, in practice most are submitted at the meeting itself, or in the four days following. Very rigid deadlines must be met and there is no time for communication with every author about editing changes made in order to improve content and style, and to satisfy space restrictions.

Every effort is made to notify an author prior to the deadline if paper is too long, or does not satisfy other restrictions on footnotes, tables, and figures. However, sometimes a manuscript is not submitted or revised on time and cannot be published. We are sorry to have to report that this occurred in one case this year. A paper is also rejected if, after reading it, we conclude that it is utterly without merit. This year we are pleased to report that no paper has been rejected on this ground.

JOHN G. RILEY  
WILMA ST. JOHN

## RICHARD T. ELY LECTURE

### Economics in Theory and Practice

By SIR ALEC CAIRNCROSS\*

Do you not know, my son, with how  
little wisdom the world is governed?

*Oxenstierna*

Let's face it. Whatever economics was in the past, it is now virtually an industry. It stretches from the building of new models by the theorists to the supply of advice, forecasts, proposals, and programs by the practitioners, and caters mainly for a market of policymakers in business and government. In the economics business, market forces work feebly, particularly at the level of theory. The competitive process derives little benefit from price adjustments, and suppliers are often remote from the market and unaware of market pressures. But the usual phenomena of growth and development are all at work: investment, economies of scale, and the interaction between technical innovation and market expansion. Some of our colleagues confine their activities to production while others occupy themselves with the business of packaging and marketing. Division of labor has made rapid progress, both horizontally and vertically. On the one hand, we have specialists in different branches of economics: macroeconomics, industrial economics, transport economics, health economics, international economics, mathematical economics, etc., etc. On the other hand, we have a lengthening chain of intermediaries between the priestly who live in clouds of theory and the lay brethren in Washington, Whitehall, and elsewhere, who do battle in the corridors of power. Where so many labor, their efforts merit scrutiny as yet another branch of economics.

I do not propose today to embark on so ambitious a task as an exposition of the economics of economics. Having spent half my working life in a succession of government departments and international bodies, I thought it best to set myself a more modest task and draw on my own experience as an intermediary in the market for economic advice. I propose to limit myself to an examination of the links between theory and practice, between the theorists who seek to trap the inner secrets of the economy in their models and the practitioners who live in a world of action where time is precious, understanding is limited, nothing is certain, and noneconomic considerations are always important and often decisive.

Action can take two forms. It may go no further than policy recommendations, or it may consist in taking policy decisions. When I speak of practitioners I shall normally have in mind those who busy themselves with what the policy should be, whether professional economists or not, rather than those who take the final decisions on policy. But I may on occasion feel obliged to refer to the difficulties of the decision taker in making use of economic advice as opposed to those of his economic advisers in formulating it.

When one looks around, theory and practice are often far apart. In many countries there is even a physical separation: the theorists remain in their universities, the practitioners in their departments of government, with little contact between the two. And since ideas circulate most freely through personal contact, the physical segregation carries with it an intellectual segregation. The thinking of advisers on policy proceeds largely in isolation from the thinking of the academics. Even in countries where there is some circulation between universities and government, and some mixing of one set of

\*St. Peters College, University of Oxford, Oxford, England.

economists with the other, there is a strong tendency for the thinking of each to stay within its own orbit, the insiders pursuing lines of thought independently of contributions from outside, and vice versa.

It is hardly surprising that there should be some divorce between theory and practice when their starting points are so different. As in medicine, engineering and other human activities, one can ask either: "what is the truth of the matter?" or: "what ought I to do?" according as one's interest is in theory or practice. An economist entering a business concern or a government department, unless consigned to the outer darkness of a research section, finds himself in an atmosphere where action takes precedence over intellectual speculation. The question at issue for the practitioner is always: "what is to be done?" That is a question which the pure theorist may decline to answer because he feels that he has no special competence to do so. He may share the view of Nassau Senior that "the conclusions of the economist, whatever their generality and truth, do not authorise him in adding a single syllable of advice."<sup>1</sup> But it is not a question that can be evaded; and presumably a training in economics is of some help in answering it.

How much help does theory provide? Sometimes the honest answer is "very little." It may elucidate, but certainly does not resolve, controversial issues of economic policy. An obvious example is the controversy in Britain in the early 1970's over the desirability of joining the Common Market, with half the academic economists signing a letter in favor, and the other half signing a letter against. Or one can point to the conflict of view between those who put their faith in monetary policy and those who regard it as a broken reed, or between the advocates and opponents of floating rates of exchange, or between those in favor of and those against a statutory incomes policy. Even when the theorists are in agreement, the issue of policy remains undecided. There is widespread agreement that in theory an expenditure tax is preferable to an income tax. But so far as I

am aware, the Finance Ministries of the world have remained unmoved. There has been no rush to change over to an expenditure tax and the only countries which did, India and Sri Lanka, gave up the experiment almost at once.

The limitations of economic theory were brought home to me when I was asked to organize a course of instruction for senior administrators who had come to Washington for six months to learn as much as possible about the kind of economic policies their countries ought to pursue. They did not want to study economic theory as such, and had indeed no time in which to master it, but were interested in the practical upshot of economic thinking and speculation about economic development. They asked quite simple questions—some of them with a familiar ring—such as: "can inflation assist or does it retard economic development? How much can one safely borrow abroad? What tax system is most likely to favor economic development?" I found, as you might expect, that economic theory was indispensable for analyzing their problems, but that it very rarely allowed one to arrive at policy conclusions with any confidence.

Later I encountered a similar group who had come to study investment appraisal and had become well versed in the theory of discounted cash flow. But investment appraisal involves a lot more than economic theory. I asked the group what rate of interest they would use on their return home. There was a long pause until one bold spirit suggested "Bank Rate." Nobody contradicted him. Nobody had other suggestions to offer.

An earlier occasion on which I asked myself to what use I would be able to put my knowledge of economic theory was when I joined the British War Cabinet Secretariat in 1940. There could be no doubt of the profound influence on policy in wartime of a comparatively small group of professional economists. And yet I never saw much use made of the more refined and esoteric parts of economic theory. I concluded, as my colleague, the late Ely Devons put it, that "in so far as economic theory is useful in enabling us to understand the real world and in help-

<sup>1</sup> Quoted by John Jewkes (1953, p. 29).

ing us to take decisions on policy, it is the simple, most elementary and, in some ways, the most obvious propositions that matter" (1961, pp. 13–14). But, as he was careful to add, before the simple propositions become part of normal processes of thinking and cease to be "kept in a separate compartment labelled 'economic theory,'" familiarity with the subject needs to advance well beyond the elementary level (pp. 25–26). Lionel Robbins said much the same when he argued that

...the most useful economic principles, when stated in their most general form, seem often mere banalities, almost an anti-climax after the formidable controversies amid which they have emerged. Yet experience seems to show that, without systematic training in the application of such platitudes, the most acute minds are liable to go astray.<sup>2</sup>

I found that two or three rather elementary economic concepts, which I had assumed would be familiar to everyone, were often not at all well understood by non-economists but were of particular value in policy formulation. Among these concepts I should include as of first-rate importance the idea of the interaction of supply, demand, and price; the concept of opportunity cost; and the marginal theory of value. Later, I concluded that it was even more important to be able to think of market forces operating within an economic system, and to recognize the coordinating function of the price mechanism. Many other elementary concepts, particularly at the macro level, were equally fundamental, but these examples are enough for purposes of illustration.

Noneconomists have rarely sorted out in their mind how supply and demand operate on market prices and have no instinctive appreciation of the virtues—indeed the indispensability—of the price mechanism. On the contrary, their bias is almost always to-

wards an organizational or political approach to economic problems. They like fixed prices because they seem to inject an element of stability and predictability. During and after World War II, when something became scarce the immediate reaction of business men or bureaucrats was nearly always in favor of control and rationing without any thought for the contribution that *some* rise in price might make to relieving or ending the shortage. The pricing of coal, for example, at the time of the nationalization of the industry in Britain in 1947, paid not the slightest regard to the chronic shortage of fuel and the danger that that shortage would arrest industrial recovery, as in the end it did. The pricing of foreign exchange, in much the same way, was divorced from market pressures and continued to be regarded by ministers as a moral or organizational issue: they believed that planning and control could do all that devaluation of the currency could do.

Of course, economists may fall into the opposite error and think that market forces, left to themselves, will always do the trick. At the end of World War II, when practically every country except the United States was running a balance of payments deficit, there were those who regarded the dollar shortage as an invention of governments that were determined to prolong the shortage by overvaluing their currencies. How far rates might have to fall and what the consequences of such a fall might be were matters rarely explored. In the early postwar years, with demobilization in progress and production well below capacity, it was not at all self-evident that a general realignment of currencies and a consequential revamping of the price structure would do much to restore balance of payments equilibrium, however necessary it might prove later on. On the contrary, there was good reason to take direct action to limit imports, encourage exports, develop alternative sources of supply and restrict the export of capital, that is, to make use of planning rather than prices.

Similarly, at the outbreak of war the necessary reallocation of manpower cannot easily be brought about by market forces alone. It might be possible in theory to work

<sup>2</sup>Quoted from an official wartime memorandum in my *Essays in Economic Management* (1971, p. 203).

through variations in the funds at the disposal of different departments and agencies, but if the government means to impose its priorities on the market it will achieve quicker and more predictable results by direct methods. Where a major upheaval is required, market forces operate slowly and blindly.

Opportunity cost is another concept that does not come naturally to the noneconomist. Few people have given thought to the inner meaning of "cost," or habitually decide on a course of action on the basis of the alternatives that might be adopted. Yet in my experience the concept is indispensable in policy analysis and lends itself to very wide applications. This is equally true of the idea of the margin: the average man thinks of averages rather than increments and often goes off on the wrong tack for this reason, particularly in relation to pricing and investment decisions.

Both concepts, however, need careful handling. Marginal theory is usually taught in terms of a single margin when in fact there are a great many. No businessman thinks of output and prices as his only variables, and even when he does, has to consider the repercussions of changing either of them over a whole series of time horizons. With opportunity cost there is a similar danger of neglecting the full range of possibilities. Lord Kaldor has recently used the concept to justify keeping open high-cost coal mines under conditions of heavy unemployment. But the logical conclusion of his line of argument is that so long as there is substantial unemployment, no firm should ever be allowed to close down and no one should ever be sacked, since it is better to have some output than none. The alternatives compared have to have regard to the full consequences, not just the immediate ones.

When I read the literature on shadow prices I have a rather similar reaction that the idea of opportunity cost can be carried too far. The notional prices corresponding to the opportunity cost of capital, labor, or foreign exchange may be enforceable on the limited sector of the economy under the government's control; but that introduces distortions between the controlled and uncontrolled sectors, which may thwart the

government's intentions. Besides, the enforcement of shadow prices that diverge widely from market prices is far from easy, even within the controlled sector. Subordinate authorities are apt to take little notice of a hypothetical test rate of discount in deciding on their investment program and do their sums on the basis of the rate they have to pay on borrowed money, diluted by any subsidies from the central government. To make a shadow rate take effect throughout the public sector, the central government is unlikely to get very far by directives unsupported by offers of capital at the shadow rate.

The biggest single advantage that economists have is their way of thinking. It comes naturally to them to think in terms of alternatives and to trace the implications of alternative lines of action within the logical framework of an economic system. They are alive to the interaction of economic forces within that system and hence to the full economic impact of policy decisions. They are not at a loss, like Prime Minister Attlee, to understand how it is that when activity is so brisk at home there should be so much trouble with the balance of payments. They do not need to be persuaded like Lord Radcliffe—perhaps the most outstanding lawyer of his day—that an enquiry into the working of the monetary system may involve a study of the working of the capital market (though I must admit that there are professional economists who even now seem to share Lord Radcliffe's view).

The importance of an adequate framework of thought was strikingly illustrated in the controversy over central economic planning after World War II. Administrators and politicians alike tended to overlook the role of the price mechanism in their enthusiasm for planning. Two of the most outstanding figures of the period, Sir Oliver Franks and Sir Stafford Cripps—one a top administrator and later Ambassador in Washington, the other a memorable Labour Chancellor of the Exchequer—published expositions of the case for central planning without any hint that there are always powerful forces at work to close any gap between supplies and requirements and that it may be well to pay

regard to these forces.<sup>3</sup> Few administrators or politicians, unless trained in economics, perceive that there can be no question of relying exclusively on government planning, or alternatively on the price mechanism, and that the real problem is always how to combine the two.

It can happen, as in wartime, that the price mechanism plays only a minor part because the government's priorities must take precedence over those of individual consumers; and in the wake of such circumstances the role of prices may be overlooked. It can also happen that economists are so mesmerized by the price mechanism that they limit their vision to the study of market forces when the phenomena of government planning merit equal attention. Just as administrators may fail to understand the workings of the price mechanism, so economists are apt to disregard organizational influences on economic activity. What goes on *inside* the firm, *inside* the government department, *inside* the Cabinet, is often left on one side. Yet it cannot make sense to pursue the study of market failure and undertake no systematic analysis of the weaknesses of alternative agencies of coordination.

To the four elementary economic concepts I have just discussed—supply and demand, opportunity costs, the margin and the economic system—I could add some familiar maxims such as “Bygones are forever bygones,” or “There is no such thing as a free lunch.” These, too, are very helpful in coping with muddled thinking in high places. They form a small but indispensable part of the economists’ stock-in-trade. Where the full range of tools is most likely to be brought into play is in economic forecasting. Here indeed the practitioner has to keep in close touch with current theory. The relationship between theory and practice in economic forecasting raises many interesting questions, since those who prepare the forecasts and are

perhaps best equipped to judge the risk of error may have little contact with those who use them and run their risks on the basis of the forecasts. But economic forecasting is much too large a subject for me to do more than touch on.

I turn instead to examine some of the reasons why economists find difficulty in bringing their theoretical apparatus to bear on practical problems. As Jacob Viner, who had plenty of experience, emphasized years ago, “the list of handicaps of the economic theorist as a participant in public policy...is discouragingly long” (1958, p. 109). Some of these handicaps arise from the practical difficulties that attend the use of economic theory in trying to work out an appropriate policy; others relate to the presentation of the policy so that it carries conviction and obtains support; others again derive from the need to marry economic with noneconomic considerations in making a policy acceptable. Let me take these in turn.

#### I. Limitations of Economic Theory

Economic theory is fundamentally an exploration of models and conceptual relationships couched in hypothetical terms and necessarily abstracts from many features of the real world. Without abstraction and simplification it would not be possible to begin thinking about economic problems. There is no option but to leave out what may seem to some people highly important. As Wicksell pointed out, it is not to be expected that economic theory should attach significance to the features of the real world according to their prominence in the eyes of the layman since “it is not the purpose of science to describe the obvious in elaborate terms” (1934, p. 19). But abstraction can be carried too far. The theorist may follow paths that lead him further and further from the real world and expose him to the danger of what one economist has called “theoretic blight” (E. R. Walker, 1943, p. 57). He may be tempted to select problems that lend themselves to sophisticated technical analysis rather than on grounds of practical importance; and become lost in admiration of the conceptual schemes he has developed without

<sup>3</sup>Sir Oliver Franks (1947); Sir Stafford Cripps’ exposition appeared anonymously in the *Economic Survey for 1947* (Cmd 7046).



regard to the unrealistic premises on which they are constructed. He may also make the common mistake of getting things the wrong way round; or leave out what really does matter or can only be left out provisionally. He may then be deceived into thinking that he understands how things work when in fact the model is misconceived. Theory, as someone once put it, can be "an organized way of going wrong with confidence." To be a useful guide it has to separate correctly what is adventitious from what is truly significant.

Economic policy, on the other hand, has to deal with practical problems and specific situations. While it is possible to develop a branch of economics bearing on these problems and situations and call it applied economics, such a branch is still part of economic theory. It still consists of a set of logically consistent propositions and abstracts from many of the circumstances that may in practice govern the policy pursued. What is to be done is never a simple corollary of theoretical conclusions.

The need for care in drawing conclusions from theory was brought home to me in Berlin in the winter of 1945-46 when I took part in a discussion between Sir Paul Chambers (later Chairman of I.C.I.) and General William H. Draper (then Economic Adviser to General Clay). Sir Paul, challenged as to the accuracy with which he had been able to forecast budgetary revenue as Director of Statistics and Intelligence in the Inland Revenue, gave us a short exposition of the theory of probability. "If you toss a penny and it comes down tails ten times in succession," he said, "that doesn't affect the probability that it will come down heads next time. The chances remain fifty-fifty." "Shall we test that?" said General Draper, producing a penny. "Will you call?" Ten times Sir Paul called heads and each time the penny came down tails. Before tossing it again, General Draper revealed that the exercise of a little sleight of hand might be affecting the behavior of the penny. It is always necessary to enquire whether the assumptions of theory are valid in the case at hand before applying it; and if the facts do not conform to theoret-

ical expectations it may be the facts that need looking into, not the theory.

Whatever the limitations of economic theory, it is very powerful stuff, more powerful the more general it comes. We certainly cannot dispense with it in trying to understand any economic system. If we enter a maze we need a thread to guide us in it and the purpose of theory is to furnish that guide. On the other hand, we cannot hope to get very far with theory alone and there are serious dangers in moving from the world of theory to the real world without regard to the difference between the two. One danger is that the theory may be obsolete. It isn't just the practical man who may become the slave of some defunct economist. Even professional economists, deeply immersed in their everyday duties in some government department, have to live on an intellectual capital that is rapidly depreciating and need an opportunity of rebuilding it in an academic environment. There may also be times when the boot is on the other foot and it is the practitioner who is alive to truths disregarded in current theory. Theory may suffer from a distortion of emphasis or a quirk of intellectual fashion that throws into prominence the wrong variables, the wrong problems or the wrong formulations of them; attention may then be diverted from the things with which theory should be occupying itself. When that happens, economic theory must be accounted not just irrelevant, but bad: for the primary purpose of theory is to assist us in posing questions, and if we are moved to ask the wrong questions theory has failed us.

The most serious problem for the practitioner is that the theorists differ, even on technical economic issues. There is no agreement on how the economy works—on what governs the level of output or employment or prices. Where the disagreements go so deep as they do nowadays it is difficult to speak with authority on technical economic issues. I need not dwell on the problems this creates in advising on policy.

And yet there are times when I wonder whether the disagreements between economic theorists, even now, go so deep as

their solidarity when confronted with the heresies which so often shape the policies of governments. To take an extreme case, we may debate whether the money supply is too great or too small: but what of governments—and there have been some—that try to do away with money altogether or come to power, like the Social Credit party, preaching that there is never enough? Or, to come nearer home, what of the comment made to me on the Radcliffe Committee by the President of the National Union of Mineworkers, one of Arthur Scargill's predecessors: "You fellows seem to worry about what the rate of interest should do. But my members don't see why there should be a rate of interest at all." Or, still on the subject of interest rates, what are we to make of Chancellors of the Exchequer who exclaim like Hugh Dalton: "You can't allow higher interest rates while resisting higher wage rates." It can sometimes be easier to reach agreement between economists on what should be done than on matters of theory.

A further difficulty facing the practitioner relates not to theory but to economic information. Economic theory has always to be mixed with a large dollop of fact before prescriptions for action can be framed; but the facts are usually obscure, disputed, seen through different eyes against a different experience of life and stretching far beyond the limited economic context within which the economist seeks to analyze them.

The theorist moreover is in control of his starting point, since he is free to make his own assumptions; but the practitioner is never quite sure where he is. As Lord Roberthall, who was Economic Adviser to the British government for fourteen years, used to say: "it's very hard to forecast where you are now." Indeed, you don't even know where you *were*. The official statisticians are busy rewriting history from the word "go"; and they don't stop. When I look back at the British balance of payments deficits in the three years after World War II, for example, I find that the figures for the current account first published added up to £1245 m., were revised by 1953 to show a total of £740 m., and continued to be revised over the next

thirty years until they dwindled to £585 m. Instead of working out at exactly the level assumed in the Washington Loan Negotiations in 1945, the cumulative deficit is now put at less than half and British capital exports over the period are consequently estimated at a total higher than was thought at the time by \$2½ b., that is, by two-thirds of what was borrowed from the United States. Another example is the way in which the U.K. monthly index of industrial production in 1964 was completely flat in the nine months up to September—a General Election was due in October—but was revised over the next two years so that it was sloping steeply upwards in official publications in 1966 and then was further revised until now it is flat again, as in 1964.

I cite these changes, which could easily be multiplied, to show that if the future is uncertain, so also is the past. I have often been intrigued to see how patiently economists apply themselves to explaining what, if later information is to be trusted, never occurred and how figures of assorted reliability are given equal treatment by those who do not live among them. The practitioner, recognizing the uncertainty of the information at his disposal, can have only a limited grasp of what is going on. He has to make the best of incomplete, inconsistent, and changeable data, relying on human judgment to derive a plausible, self-consistent picture of the existing situation. He is quite likely to find, as I have found, that the best way to reconcile the data is to begin by making a forecast of the future as a way of deciding on the underlying trends and then work backwards to a consequential interpretation of the present. The judgment he makes—as in the examples I have cited—may be crucial to the choice of policy. If for instance, you think the economy is stuck, you opt for policies very different from those appropriate to a rip-roaring expansion.

A further difficulty is that the economy never works in quite the same way for very long. You may feel confident that you can explain how it worked in the recent past and set your conclusions down in equations with all the coefficients, lags, etc., carefully esti-

mated. But, as Keynes put it, human behavior is not "homogeneous through time."<sup>4</sup> Whether you realize it or not, you are always working with relationships that are obsolescent without knowing just how obsolescent they are. One day you can count on people spending more when prices go up; then you find them spending less. One day the unemployment figures go up when the vacancy figures come down; then they both go up together. It is always necessary to be on the look out for some departure from normal patterns and pay attention to straws in the wind. They may reveal, earlier than any statistics, new forces at work or a strengthening of existing forces. Analysis of these forces has to be coupled with a good eye for straws.

Then there is the limitation imposed by the need to be specific: in particular, to deal in specific magnitudes and at specific points in time. Many of the more important generalizations in economics make no reference to magnitudes or time. They may be of assistance to a government that wants to know in which direction it should be operating; but they do not, in their general form, offer much help to a government wanting to know how far to go.

For example, it may be possible on general grounds to indicate that the government should be thinking in terms of increasing taxation. But the question that has operational significance is, how much should the increase be? This requires immersion in a mass of statistical detail and the working out of far more definite views of the functioning of the economy than found their way into the traditional textbooks in economics some years ago.

Then there is the content of the tax package. What *taxes* should be increased? What effect will the increases have? What other action, if any, should accompany the increases in tax or be contemplated for introduction later?

Another set of issues relates to timing. When should the government act? When will

it be possible to judge whether the action has been effective? Is it likely to be necessary to take further action later?

It takes time to become aware of changes in the situation, to size up the strength of the forces at work, to prepare the appropriate response. One cannot wait for certainty, but it is also a mistake to act prematurely when the diagnosis may prove to be quite wrong. Delay may be inescapable. After the devaluation of sterling in November 1967, there was a great burst of consumer spending and a clamor for early action to restrain it. The right time to act was of course in November, but when that opportunity was missed it was not easy, for technical reasons, to redeem the error by imposing additional taxation in the weeks immediately before Christmas. In January it seemed better to put all possible effort into a battle for lower government expenditure and by the end of the month the budget was already in sight only a few weeks away. So although the need for action was not in dispute, it was four months before a suitable package of measures could be introduced.

Another source of difficulty is that many of the questions on which advice is sought from economists have very little to do with conventional economics. Cabinet ministers, I found, don't ask the questions you are ready to answer. They want to know how people will react, both to events and to their policies. Will there be a strike or won't there? Will the rate of exchange weaken or strengthen? Will it be possible to get backing for this or that line of policy?

I concluded that attitudes were just as important as prices and that economic policy had to embrace efforts to change attitudes, not just efforts to make better use of market forces. Just as economic events and policies may have their biggest impact outside the functioning of the economy—as world depression could clear the way for Hitler—so of the most effective levers of economic policy sometimes bypass the market operate on confidence and opinion, expectations and attitudes. In the same way as economists so often neglect goodwill in discussions of industrial economics, so they tend to neglect the prestige, credit, standing,

<sup>4</sup> Keynes to Harrod, *Collected Writings* (1973, pp. 296–97), quoted by Bernard Corry (1978, pp. 5–6).

authority—call it what you will—of governments and the ways in which morale and endeavor are affected by factors other than pay.

## II. Presentational Difficulties

Let me turn next to presentation. This raises problems at two levels, that of the theorist and that of the practitioner.

Theorists may confine themselves to the business of producing theories without much regard to the market for them. But in applying economics in practice, it is impossible to overlook the importance of the consumer. This is obviously true in the short-run sense that one has to have regard to the chances that any attention will be paid to suggested lines of action by those who have it in their power to act on them. It is true also, in a much wider sense: that those parts of economic theory that do not supply useful answers tend to receive little attention in business or government, while those that purport to throw light on practical problems, and point in the direction of specific ways of dealing with them, command respect and interest.

Practitioners face a rather different problem of presentation. Governments are almost as much concerned about what to say and how to say it as about what to do. Indeed, what they say may have more effect on the markets than what they do. They may be given credit for cutting public expenditure by simply announcing that that is their intention even when, as in the first four years of Mrs. Thatcher, it continues to increase. Similarly they may be given credit for mastering inflation when all they have done is to contribute to an international depression that brings down import prices. The public reacts to the declared aims of government as presented in speeches, often without close enquiry into the success with which these aims are pursued. This being so, economists can neither ignore how policies are presented nor how market opinion may narrow the scope for government action. Against the extra leverage that skillful presentation of policy may provide must be set the danger that the government may become the prisoner of

market opinion, forced to conform to the role assigned to it by that opinion, and so transmuting into rational expectations what would otherwise have no rational foundation.

The issue of presentation is obviously highly important when any major change of policy takes place. If, for example, a more restrictive policy is proposed involving higher taxes, the Minister of Finance needs to see the case presented in persuasive terms so that the government, in turn, can be persuaded and the new policy defended in public debate. There is always a question how the higher taxes can be presented with the minimum damage to the credit and authority of the government and its capacity to carry through the rest of its program. What is to be said and how is it to be said? The handling and presentation of the decision is part of the decision itself and cannot be dismissed as irrelevant to it. It is partly because this is so that it becomes difficult to find a use for those parts of economic theory that are not easily translated into simple language.

Taxation provides many illustrations of the problem of presentation. I can remember Chief Festus of Nigeria recounting how he had to withdraw a tax on cosmetics because, as he explained, holding up a large, pudgy hand, "I burnt my fingers." In Britain the Selective Employment Tax introduced in 1965 was withdrawn six years later, in part at least because the refined economic logic by which it was justified did not make sense to the general public. Or take corporation tax. Economists might agree that there is no strong case in theory to have a corporation tax at all. But a proposal to abolish the tax would certainly be laughed out of court by politicians and would be unintelligible to the general public.

In stressing presentation and acceptability, I should not want to be interpreted as defending mere sycophancy and time-serving, automatic approval of any act of policy that is likely to gain popular approval and command a Parliamentary majority. Neither Parliament nor the public has any prerogative of wisdom in economic affairs, whatever democratic theory may imply, and the test of sound policy can never be made on accepta-

bility alone. On the contrary, the economist is wise to be on his guard, as Marshall emphasized, when his views are popular and all men speak well of him (cited by Pigou, 1925, p. 89). He owes it to his profession to speak up for what he thinks right, to denounce policies that he thinks mistaken and to try to persuade those in power of the dangers they run if his advice is neglected. But if he wishes to be heard, he has to learn when to keep his peace and when to press his point. There are times when policies have to be ruled out because the political leadership required for their adoption simply does not exist; and when indeed the policies that seem right to the economist in his study might provoke adverse reactions, of which he has taken little or no account, but would make nonsense of the policies. There are other times when new ideas could fill a political vacuum, and what was previously unacceptable can be taken down from the shelf and put on sale.

In practice, political choices are rarely a matter of good and bad, black and white. They usually turn on a balance of considerations among which economic factors are not decisive. I don't know what undergraduates make of the questions they are asked to decide in three-quarters of an hour in final examinations. But if they have difficulty in coming to firm conclusions they are in good company. One can make a case—and generally quite a respectable case—for a variety of economic policies at any point in time and argument is unlikely to destroy every case but one and leave the surviving case as indisputably "right." Economists do notoriously disagree. So what they have to square with their conscience is usually not failure to demonstrate the error of some politician's ways but failure to offer the right degree of resistance, to do battle with the right degree of conviction, to use what Lord Roberthall once designated "the right tone of voice." Like the lawyer, the economist comes to see the case that can be made for and against, and loses the campaigning spirit with which he set out. He has to be forever pointing out that things are not quite so simple as politicians suppose, forever dwelling on the hidden snags. Policies cease to be right or wrong,

but just better or worse, and often only marginally so. The occupational disease that he has to fight is not time serving but atrophy of conviction and the sense of commitment.

### III. Noneconomic Factors in Economic Policy

I come next to the implications of the obvious fact that the policies of governments are by definition a political matter. If you are considering what governments should do, you can hardly avoid taking account of what sort of government you have, and how much government you want. It makes quite a difference whether you have been brought up to regard the government as Santa Claus, Stalin, or a dog fight. One government may be benevolent, another dictatorial, a third incapable of making up its mind. All of them have the failings of their human components, ministerial and bureaucratic.

Governments are political animals, moved by political considerations. They have to ask themselves what they *can* do and this may rule out many otherwise attractive lines of action. There are commitments by which they are bound—to other governments, to particular interests, to the party supporting them in office. They hesitate to fly in the face of prevalent attitudes and opinions. They are more conscious of immediate pressures and short-term considerations than of what is desirable in the long run and usually prefer to put off the evil day. Even when they are anxious to do the "right" thing, as a surprising number are, or when they give priority to long-term objectives over short, they tend to do so with an obstinacy fatal to their hopes: either because they lack understanding of the appropriate sticking-points, or because they hesitate to give ground for fear of unsettling opinion and losing the support they need. One of the most difficult problems in policymaking is to know how far to persist and when to bend. Overcommitment can be worse than opportunism.

In any event, the economist has to recognize that policy does not take shape in a vacuum but within a machine that has several well-defined organizational characteristics with which he would do well to become acquainted. Government is not a simple opti-

mizing activity that can be reduced to a second differential in a mathematical equation. It is more likely to be a collection of bald-headed and somewhat bewildered men sitting round a table, harassed and short of time, full of doubts and dogmatism, with all the strengths and failings of successful politicians. Such men may survive for a long time without any policy at all except in the form of a series of specific responses to matters forced on their attention and calling for immediate decision.

If, therefore, the economist wants to influence policy and asks where policy is formed, the answer may be either anywhere or nowhere. It is not unknown for political theorists studying a government department to come to the conclusion that no intelligible answer can be given to the question: "who forms policy?" A succession of battles on a succession of issues may rage within or between departments, involving different groups at different times, and there may be no consistency in the outcome of their debates except what is imposed in ignorance by some later historian. Or decisions may be taken low down in the hierarchy by someone who is unaware that he has taken any decision at all (such as the decision to do nothing); and although the matter may be fought out at increasingly exalted levels until it reaches the Cabinet, ministers may have no option but to accept the inevitable, even if so little disposed to recognize their own impotence that they go through the charade of further debate and carefully minuted decision. It is one of the curiosities of government how frequently what is plainly due to the force of events is attributed to free and deliberate choice.

This is not to say that policy itself is a hallucination and not worth bothering about. What governments do can hardly be discussed in such a ludicrous fashion. The point is rather that one has to understand the scope for policy, the times at which it may be influenced, and the pressures that govern it. Similarly, one has to have some awareness of the bureaucratic atmosphere within which economic problems arise and have to be tackled. That atmosphere is somewhat different from the comparative calm of university

life. Many years ago I described how "the various divisions in many government departments (were) loosely geared together, uncertain of the limits of their responsibilities, losing and gaining staff almost every week, themselves dissolving from time to time into new divisions or subdivisions, and facing an avalanche of fresh problems on which to advise, fresh cases to decide, and fresh policies to apply."

No doubt that exaggerates a little; but it brings out some of the features of life with Leviathan that an academic economist might overlook. These features condition the way in which economic theory impinges on policy and limit in particular the chances of drawing on highly complex bits of theory.

Allowance has to be made next for the political setting: the need, if one is in business, to guess what the government will do next, or, if one is in a government department, what is likely to prove feasible and acceptable to a government wishing to stay in office. A wise decision on what should be done cannot be based on economic reasoning or models that pay no regard to the distribution of political power, the frame of mind of the public, or the political ambitions and anxieties of the party in office.

Suppose, for example, that one thinks, like one of my distinguished Cambridge colleagues, that the economic situation calls for the use of import restrictions. One may begin by setting out the economic arguments. Then one has to reflect on the political situation. If on January 1, 1973, Britain has just joined the European Economic Community, one has to ask whether it makes sense to urge ministers *a week later* to introduce import restrictions that will fall heavily on imports from Common Market countries. If in June 1975 a referendum is to be held on continued membership of the Community and the Cabinet is split down the middle on the issue, one has again to ask if it makes sense to press the Chancellor, just ahead of the referendum, to budget in April for import restrictions, especially if the identical remedy was appropriate two years earlier in very different circumstances. If the advice is accepted and an international row brings on a run on the pound, how is the Chancellor to explain to

the IMF that he acted in the interests of greater stability in the exchange rate, and how is he to put it to his continental colleagues—most of them struggling with heavier unemployment than Britain—that he felt compelled by the intolerable level of unemployment to set aside his treaty obligations.

It is not only the organizational and political setting that is important. Economic problems have also to be seen in their institutional setting. It is (or should be) impossible to discuss monetary policy without regard to the kind of banking system and methods of credit control in operation, just as it is or should be impossible to discuss wage theory without regard to the way in which wage bargains are struck and bringing in various kinds of legislation affecting bargaining power (for example, in relation to minimum wages, the powers and practices of trade unions and employers' associations, redundancy, labor mobility, and so on).

Frank Knight in his latter days used to agonize over the futility of being an economist. He doubted whether society would ever take advantage of anything he had to contribute to the solution of its problems. Others like Max Planck have turned away from economics because of its appalling complexity. Others again have given up in despair of arriving at finality: they are repelled by the inconclusiveness of the subject—what Wicksell called “the permanent state of war” (1958, p. 52) between diametrically opposed views neither of which is ever vanquished or disappears from the field as would happen with the natural sciences. There is no received body of doctrine—only a “technique of thinking.”

In spite of what I have said about the limitations of economic theory as a guide to policy, the contribution it can make seems to me none the less invaluable. Any doubts on that score are soon quelled by life among noneconomists in positions of power. Moreover, the very inconclusiveness of economics has its value as a preparation for the world of affairs where the same inconclusiveness rules. In government and business there is rarely a conclusive answer; instead there is an equally enduring “state of war.” The evi-

dence on which an answer might be reached, even on matters of fact, tends also to be inconclusive since there is rarely any finality in the statistical data that purport to summarize the facts. It is necessary to decide between alternatives in the light of uncertain and often contradictory evidence. The decision, it is true, rarely turns exclusively on economic considerations. But it is a great advantage to be able to assess the force of these considerations, just as it is also a great advantage to be able to test the data with the kind of insight into the underlying relationships at work that economic theory engenders.

In the application of economics to practical problems, that kind of insight needs to be reinforced by imagination and accurate observation. Imagination is kindled by good theory but is powerless or mischievous if fed with inadequate or inaccurate information. In the social sciences there is no substitute for getting the facts right, and observation ranks at least as high as logic. Most theoreticians tend to treat far too lightly the difficulty of obtaining and presenting the information necessary to a sound decision. If you want to understand how the economy works, you need to have an eye for the information that matters; and since the unexpected keeps happening you need very up-to-date information. An economist like Keynes may owe his reputation to his originality as a theorist; but in my judgment he stands out from the other economists of his time at least as much for his flair in picking on significant statistics, often from relatively obscure sources before anybody else, and piecing them together by conjectural arithmetic to reveal a danger not then fully appreciated. Other economists of the first rank commonly have a similar power to startle with unfamiliar figures that give a new perspective to events.

Those who have done their homework thoroughly, and have mastered every scrap of information likely to be of assistance will be of little use, however, without the imagination to conceive of alternative policies and visualize the reasons why they may not work as expected. They may fail to make use of available information because they do not appreciate its relevance and overlook or mis-

construe important relationships. For example, price control is obviously not enough by itself to remove the danger of inflation in wartime. But it required considerable imagination in World War II to invent three new devices for that purpose: postwar credits (an acceptable form of forced saving); points rationing (the circulation of a new currency to be used exclusively for the purchase of rationed goods); and subsidies to stabilize the prices of key commodities, making up a kind of iron ration. All these were expedients, not intended to last indefinitely, but they did contribute to a general stabilization of incomes and prices.

#### IV. Conclusions

It has been part of my theme that economics has more to offer by way of analysis than prescription. So it is hardly surprising that I should have few proposals for improving the state of affairs I have described. I have three rather modest suggestions.

The first can be put in a word: circulate. The practitioners need to mix with the theorists and vice versa. More than that—the practitioners need to be given a chance to catch up with theoretical developments by release from time to time from their duties. They should be offered sabbatical leave, or enabled to attend conferences or at the very least given time to read the journals. They also need encouragement and opportunities to make their own contributions to current theoretical controversy. Conversely, the academics need a modicum of experience of policy formulation. A spell in government or business can do wonders in changing the outlook of a theorist on the best way of spending his time, on the choice of problems to study, and on the limits within which action can be taken. In some countries, however, including my own, it has become more difficult to move in and out of government. Twenty years ago a remarkably high proportion of top British economists had had experience of service in government. Today there is very little circulation. That seems to me a step in the wrong direction.

Secondly I think we need to revalue and upgrade the work of intermediaries between

the profession and the public. Financial journalism, for example, is an increasingly demanding skill and has become both more sophisticated and professional and more influential since the war. The press also carries articles by professional economists, and a number of specialized publications reprint (or commission) articles by them that help to illuminate current issues. But the mass media are largely untouched by this trend. It may be that nothing can be done about this. But there does seem to me great scope for those economists who have a gift for conveying the thinking of the profession, with all its doubts and dissensions, to the man in the street.

Finally, don't let us be overwhelmed by our disagreements: we have also plenty to agree about. As I have tried to show, it is often the most elementary propositions in economics, on which we all agree, that matter for practical purposes. Similarly, we should not underrate the value of the habits of mind that are nourished by economic analysis, even if they yield no common program of action. Where we continue to disagree, let us try to understand and narrow our differences, remembering always that we have a duty to our fellow citizens to offer them the best advice we can.

#### REFERENCES

- Cairncross, Alec, *Essays in Economic Management*, London: George Allen and Unwin, 1971.
- Corry, Bernard, "Keynes in the History of Economic Thought," in *Keynes and Laissez-Faire*, London: Macmillan, 1978.
- Devons, Ely, "Applied Economics—the Application of What?," in *Essays in Economics*, London: George Allen and Unwin, 1961.
- Franks, Sir Oliver, *Central Planning and Control in War and Peace*, London: London School of Economics, 1947.
- Jewkes, John, "The Economist and Public Policy," *Lloyds Bank Review*, April 1953, 28, 18–32.
- Keynes, John Maynard, *The General Theory and After: Part II, Defence and Development*, Collected Writings, Vol. XIV, London: Macmillan, 1973.



## Post-Entry Competition in the Plain Paper Copier Market

By TIMOTHY F. BRESNAHAN\*

This paper reviews events in the plain paper copier (PPC) market immediately after Xerox's monopoly ended. Xerox's behavior and that of a flood of PPC entrants are viewed through the lens of recent advances in the theory of entry and entry deterrence. The events of the early post-entry period also cast some interesting light on the theory of technological competition.

The modern theory of entry deterrence rests on a simple, if not obvious, proposition. The (socially) worst industry performance is after entry, the more monopolies there will be, since the interests of the entrant and society are opposed once entry has occurred. A series of papers have considered endogenous changes in the conditions of competition-monopolists who make their industry more competitive (conditional on entry) in order to deter potential entrants.<sup>1</sup> There are two distinct steps in the entry-deterrence argument. First, it must be possible for events during the monopoly period to affect post-entry competition. Some intertemporal complication must be present, either in costs or in firm-specific demand, if the state of the industry at the time of entry is to form important "initial conditions" for competition. Second, the monopolist must find it profitable to manipulate the initial conditions by some pre-entry action.

The general theoretical questions of entry and deterrence have been cast in quite spe-

cific terms for the problem of technological competition. One view emphasizes the "Arrow effect" (Kenneth Arrow, 1962; Jennifer Reinganum, 1983; Drew Fudenberg and Jean Tirole). Because any innovation destroys some of the rents to older products and processes, incumbent monopolists have a smaller incentive to innovate than potential entrants. Another view (Richard Schmalensee, 1983; Richard Gilbert and David Newbery, 1982) points out that the incumbent's losses from entrant's innovation create a motive for preemptive R&D, product introduction, or patenting. If incumbents are leaders and entrants followers, the second view will hold independent of technology. Note that the difference is over the profitability of entry deterring strategies. In both views, the presence of valuable assets like patents or secrets provides the necessary intertemporal link.

Events in the PPC market during the time of Xerox's monopoly did have a substantial impact on the nature of competition in the early postentry period. An Arrow effect is evident, as are other equilibrium explanations of Xerox's rapid decline. The alternative explanation that Xerox was "fat" is also considered below.

### I. The Late Monopoly in PPCs

Xerox the monopolist did three things which form the basis for my investigation: it price discriminated, priced far from costs, and patented every imaginable feature of the copier technology. The early entrants (in the early 1970's) and the effects of the FTC consent decree requiring Xerox to license its patents to all comers (in 1975) were substantially affected by the initial conditions at the beginning of the competitive period.

<sup>†</sup>*Discussants:* Richard J. Gilbert, University of California-Berkeley; Paul David, Stanford University; Kenneth Judd, Northwestern University.

\*Assistant Professor of Economics, Stanford University, Stanford, CA 94305.

<sup>1</sup>See A. Michael Spence (1977); Avinash Dixit (1979): See also papers reviewed by Drew Fudenberg and Jean Tirole (1984).

Xerox used a long list of price discrimination devices.<sup>2</sup> Among other devices, Xerox based rental prices on the number of machines a customer used, whether different models were rented by the same customer, by the number of copies per month, and by the number of copies per original. Implementation of this price discrimination scheme required a "lease-only" policy. This is the first of the initial conditions of the period of competition. Xerox had an extremely large rental fleet of copiers at the time of entry. As a result, the capital loss on existing copiers due to competitive price falls would be born by Xerox, not by old customers.

The second initial condition was the result of the umbrella provided by Xerox's price-cost margin. In the small (low-volume) copier market segment, "coated paper" copiers held on to a substantial market share despite their great inferiority to *PPCs*. In the high-volume copier/duplicator market segment, photography-based methods remained common despite their cost disadvantage relative to electrostatic copying ("xerography"). Substantial product-development and distribution-network quasi rents in those industries would be destroyed if *PPC* prices fell.

When IBM and Litton entered the *PPC* market in 1972, Xerox sued to block entry under literally hundreds of patents. IBM had spent millions to "invent around" Xerox's major patents—with 25 percent of the budget going for patent counsel, not *R&D*. Later entrants depended on antitrust countersuits rather than plans to defend the patent suits, as did Litton. Over the same time period, the FTC brought another antitrust action. Over the objections of existing entrants (especially SCM), the FTC forced Xerox to license its patents to *all* entrants at nominal costs.<sup>3</sup>

<sup>2</sup>This discussion, and much of this paper, is in deep debt to staff work done at the FTC in connection with the Xerox case. Comments by R. Gilbert, F. M. Scherer, and several members of the FTC staff on my larger paper (1985) on which this one is based were very helpful. Factual assertions will not be documented in this paper; citations and more detailed evidence are available in my earlier paper.

<sup>3</sup>Existing entrants proposed an arrangement in which they and Xerox would exchange patents.

This was the third initial condition: potential entrants could be reasonably certain as of 1974 that there would be free access to *PPC* technology.

## II. The Transition to Competition I: Prices and Market Shares

Plain paper copiers are complex, highly technical products. Models vary in speed, ability to collate, needed warmup time, and reliability. Firms vary in the degree of non-hardware service they bundle to their machines.

Table 1 shows rates of change of prices indexes for *PPCs* calculated on the basis of standard lease contracts for those machines which do not change (hardware) features between years. These indexes are biased upward when new machines are introduced: in general we would expect the price fall from an existing machine to be less than proportional to the (implicit) price fall from a new machine. Yet they show an extremely rapid fall in the price index immediately following entry. Xerox's prices lag entrants' in this period. (1977 was a major new-product year for Xerox, so the "Xerox" column would show a large drop if it could be calculated.) After 1978, price indexes for Xerox (not shown) move quite closely with those for other firms.

The simultaneous existence of rental and purchase markets complicates the calculation of market shares. Two obvious definitions are possible: share in "installed base," total machines in use, and in "net new placements," current sales plus new rentals minus returns of old rental machines. Xerox's share on both definitions was 100 percent in early 1972. Its share of new placements fell to 58.5 in 1973, 43.2 in 1974, 14.1 in 1975, and 13.7 in 1976. It's share rebounded to 44.5 in 1977 and has remained since then in the 40–50 percent range. In installed base, Xerox's share declined steadily to 55 percent in 1977 and declined more slowly thereafter to a current range of about 45 percent.

The curious thing about this particular transition is the dip in Xerox's market share in net new placements. Two points are worth noting. The first is that during the transition

TABLE 1—PPC PRICE INDEXES<sup>a</sup>

	Xerox	All Firms
1973	-6.9	-8.4
1974	-12.0	-11.8
1975	-2.5	-5.9
1976	-8.2	-2.7
1977	<sup>b</sup>	-7.1 <sup>c</sup>
1978	-6.1	-6.3

<sup>a</sup>Average percent change in real rental price since previous year.

<sup>b</sup>Not calculable: insufficient comparable contracts.

<sup>c</sup>Does not include Xerox Corporation.

period, Xerox did not function as a leader relative to other firms. One could either use the language of the contemporaneous trade press—Xerox “passive” and “fat”—or note the incentive effects of the Xerox rental fleet left over from the monopoly period. In the transition period, each addition to the installed base lowers price, and it is Xerox that takes the largest capital loss on that price fall, since Xerox is the owner of the bulk of the inframarginal units. (The intuition here is that of Cournot.) Thus Xerox is at a considerable strategic disadvantage. After Xerox's installed base declines to what now appears to be Xerox's steady-state market share, this comparative disadvantage is wiped out.

The second thing to note about the entry period is that the transition took almost five years from initial entry and three years from near certainty that the FTC would allow free entry. The adjustment costs for a *PPC* entrant are nontrivial: a distribution network must be set up, a machine must be designed, and so on. It is hard to imagine, however, that these adjustment costs are particularly large by the standards of manufactures generally. This long adjustment period suggests that the entry deterrence theories may be correct in those industries, unlike *PPCs*, where initial conditions tend to give incumbents a strategic advantage.

Theories of entry deterrence and of the persistence of monopoly have naturally emphasized those intertemporal cost and demand relationships which give strategic advantages to incumbents. These are relationships of *complementarity* over time. If capital is long lived, higher production dur-

ing the monopoly period lowers post-entry costs. If users of the monopoly product invest in knowledge of how to use it, their demand curves for that specific brand will be shifted out post entry. By contrast, substitutability over time leads to incumbents' strategic disadvantage, as in the *PPC* case.<sup>4</sup> Since photocopiers are a durable good, their demand is characterized by intertemporal substitution, to Xerox's disadvantage.

Since a large fraction of modern monopolies are producers' durables, there seems to be a presumption that the most common strategic effect is incumbents' disadvantage. This presumption is too hasty, however. At least in electronics, monopoly products are frequently associated with substantial investment in information goods by downstream firms. For example, mainframe computers require downstream investment in software and in human capital, which can lead to a substantial complementarity over time in single-brand demand. This may well give IBM a strategic advantage, contributing to its continuing dominance.

### III. The Transition to Competition II: Innovation

The transition period saw a great deal of innovative activity from entrants and Xerox. It is possible to say something about the impact of competition on the direction of inventive activity based on this experience, although little information is provided about the rate of activity.

The new-product choices of firms whose rents were destroyed by the increase in *PPC* competition are illuminating. We can divide these firms into three groups: producers of coated paper copiers (*CPCs*), producers of photoduplicators, and Xerox, a full-line *PPC* manufacturer. The producers of *CPCs*, having substantial marketing and distribution expertise in the small-volume copier segment, would seem to be a natural group of

<sup>4</sup>This discussion presumes that the incumbents' products and the entrants' products are in a relationship of “strategic substitutability” post entry. See Jeremy Bulow et al. (1984) and Fudenberg and Tirole for a fuller discussion of the strategic issues.

entrants into the *PPC* business. Most did enter, but not into the small-volume segment: in 1976 a group of these firms (SCM, Dennison, AM, A.B. Dick, and Royal) was offering more than three times as many high-speed as low-speed copiers. There was very substantial entry into the low-volume segment, overwhelmingly by firms not previously in the market (Savin, Ricoh, etc.). With the exception of Kodak, which continued to offer differentiated high-volume copiers with all available technologies, photocopier manufacturers did not enter the *PPC* market. Xerox introduced new products in all segments. As a regularity, firms that had a choice chose to enter product segments where higher rates of inventive activity would destroy others' rents, not their own.<sup>5</sup>

As in the last section, we see that initial conditions matter for competition, and in the way suggested by the theory. It is again the incumbent's strategic disadvantage that shows most clearly in the *PPC* case. The logic of the entry deterrence theories is vindicated, but there is no presumption that the demand and technological conditions leading to the persistence of monopoly are common.

Verifying the generality of the *PPC* experience is no easy task. One might be tempted simply to ask whether monopolies and near-monopolies persist in high-technology areas. But all theories will imply that, in a suitably uncertain world, some monopolies will persist and others will not. No preemption theorist feels the results are weakened by RCA's failure (during the 1950's) to preemptively invent processes for producing integrated circuits, nor can any believer in the Arrow effect rule out the theoretical possibility that IBM simply built a better computer than Burroughs (in the 1960's).

<sup>5</sup>Though the immediate post-entry period saw a very rapid increase in the overall level of *PPC* innovative activity, it is not possible to draw any conclusions about the relationship between competition and innovation. It would be extremely difficult to distinguish the effects of important changes in technological opportunity, such as those resulting from the invention of the micro-processor, from changes in the conditions of competition.

#### IV. What is "Fat?"

The trade press of the transition period made much of the idea that Xerox was "losing" because it was fat, language not particularly attractive to economists. There is hard evidence of this fat, however. In the monopoly period, using very successful price discrimination devices, Xerox had price-average cost margins of around 10 percent. From 1972-76, a period with no important advances in *PPC* manufacturing process, prices of standard contracts fell by 30 percent and Xerox's price-discrimination practices ended. Other firms found it profitable to operate at the new prices. Xerox's accounting average cost must have overstated true marginal cost by at least 20 percent.

The trade press also offers a view of the differences among companies' technological tendencies which may imply a theory of where the money went. In the monopoly period, Xerox was a highly innovative firm. But the innovations were characterized by being "in the copier." These were purely technical advances in the quality of reproduction, as well as such features as two-sided copying, reduction, etc. The innovations by entrants include some of a very different character, those oriented toward the "user interface." Entrants introduced document feed devices (Kodak 150), automated many of technical features like two-sided copying (IBM III), and invented whole new product markets like the "convenience copier" (Savin 750.) One of the natural claimants for the rents in a high-technology activity is the engineering staff, not capital. Since capital in fact owns the rents, this requires an explanation of a failure of monitoring. Though taking the rents in the form of cash may be easily monitored, taking the rents in the form of a persistent bias away from the highly commercial activity toward activity of more purely engineering interest may be much more difficult to monitor.<sup>6</sup>

<sup>6</sup>This theory is complementary to Michael Salinger's (1984) observation that union labor may claim the rents.

### V. Conclusion

This attempt at an industry case study may well appear clumsy to experienced practitioners of the art. I note only my surprise at how helpful recent advances in economic theory can be in understanding real competition.

### REFERENCES

- Arrow, Kenneth, "Economic Welfare and the Allocation of Resources to Innovation," in R. Nelson, ed., *The Rate and Direction of Inventive Activity*, Universities-National Bureau Conference Series, No. 14, New York: Arno Press, 1962.
- Bresnahan, Timothy, "Impact of the FTC Consent Decree in the Plain Paper Copier Market," mimeo., Federal Trade Commission, 1985.
- Bulow, Jeremy, Klemperer, P. and Geanakoplos, J., "Multimarket Oligopoly," mimeo., Stanford University 1984.
- Dixit, Avinash, "A Model of Duopoly Suggesting a Theory of Entry Barriers," *Bell Journal of Economics*, Spring 1979, 10, 20-32.
- Gilbert, Richard and Newbery, David, "Preemptive Patenting and the Persistence of Monopoly," *American Economic Review*, June 1982, 72, 514-26.
- Fudenberg, Drew and Tirole, Jean, "The Fat-Cat Effect, the Puppy-Dog Ploy, and the Lean and Hungry Look," *American Economic Review Proceedings*, May 1984, 74, 361-66.
- Reinganum, Jennifer, "Uncertain Innovation and the Persistence of Monopoly," *American Economic Review*, June 1983, 73, 741-48.
- Salinger, Michael, "Tobin's  $q$ , Unionization, and the Concentration-Profit Relationship," *Rand Journal of Economics*, Summer 1984, 15, 159-70.
- Schmalensee, Richard, "Advertising and Entry Deterrence: An Exploratory Model," *Journal of Political Economy*, August 1983, 90, 636-53.
- Spence, A. Michael, "Entry, Capacity, Investment and Oligopolistic Pricing," *Bell Journal of Economics*, Autumn 1977, 8, 534-44.

# **R&D Appropriability, Opportunity, and Market Structure: New Evidence on Some Schumpeterian Hypotheses**

By RICHARD C. LEVIN, WESLEY M. COHEN, AND DAVID C. MOWERY\*

One of the largest bodies of literature in the field of industrial organization is devoted to the interpretation and testing of several hypotheses advanced by Joseph Schumpeter (1950) concerning innovation and industrial market structure. One set of hypotheses focuses on the role of firm size as a determinant of *R&D* spending and the rate of technological advance. Another set focuses on the effect of market concentration on *R&D* and technological advance. In this paper, we reexamine the latter set of hypotheses at the industry level, using new data on *R&D* appropriability and technological opportunity collected by Levin et al. (1984) in a survey of *R&D* executives in 130 industries.

## **I. Motivation**

Despite the voluminous literature, theory yields ambiguous predictions about the effects of product market concentration on *R&D* spending and on innovative performance. For example, it is often argued that firms in concentrated markets can more easily appropriate the returns from their *R&D* investments. On the other hand, it is argued that gains from innovation at the margin are larger for competitive firms than for monopolists, but this argument neglects systematic differences in the probability of imitation or in costs of adjustment. Schumpeter himself emphasized that concentration reduces market uncertainty and provides the

cash flow required to engage in costly and risky *R&D* on an efficient scale. Others have argued that insulation from competitive pressures breeds bureaucratic inertia and discourages innovation. Still others have used a combination of arguments to rationalize the "inverted-U" relationship frequently observed in the empirical literature, whereby innovative effort or innovative output first increases with concentration and then decreases.

One reason for skepticism about the existence of a direct effect of concentration on innovation is that theoretical arguments justifying such an effect tend to appeal to more fundamental technological or institutional conditions. The empirical literature lends some support to the view that innovation depends on more fundamental conditions. F. M. Scherer (1967) found that the statistical significance of concentration was substantially diminished when a vector of dummy variables categorizing industries by the nature of their technology (chemical, electrical, mechanical, and traditional) was added to the regression. Scherer and others have interpreted these categorical variables as proxies for technological opportunity. Technology classes differ in more than opportunities for technical advance, however; they also differ in the inherent ease of imitation and in the strength of patent protection. The technology class variables used in the literature are thus best interpreted as proxies for both opportunity and appropriability.

Concentration may also proxy for a broader range of industry-specific effects, as suggested in John Scott's (1984) study of *R&D* spending at the business unit level. Using Federal Trade Commission (FTC) data, Scott first replicated the standard inverted-U result, but when fixed effects were added for two-digit industries, as well as for

\*Levin: Professor of Economics and Management, Yale University, New Haven, CT 06520-1972; Cohen and Mowery: Assistant Professors of Economics and Social Science, Carnegie-Mellon University, Pittsburgh, PA 15213. We acknowledge the support of the Division of Policy Research and Analysis, National Science Foundation, and the valuable assistance of Margaret Blair and Andrea Shepard.

companies, the *t*-statistics on the concentration terms dropped by an order of magnitude.

These results suggest that industrial concentration may have no independent significance as a determinant of innovative effort or innovative output. To explore this issue, we attempt to control for systematic interindustry differences in appropriability and opportunity, using data from the Levin et al. survey. We also take account of the possible effects of *R&D* and innovation on market concentration, recognizing that Schumpeter's insights about the role of innovation in determining market structure may be more fundamental than his widely tested hypotheses concerning the feedback from market structure to innovation incentives.

We estimate equations for both *R&D* and innovative output. In the best of worlds, these equations would be embedded in a structural model and derived rigorously. This is the approach taken in related work by Levin and Peter Reiss (1984). We believe nonetheless that there is ample justification for the kind of "pre-theoretical" statistical investigation reported here. Theorists often begin with the "stylized facts." We offer this exploratory data analysis with the hope of providing to other researchers a more accurate assessment of the stylized facts.

## II. Empirical Results

We take as our unit of observation the line of business (*LB*) as defined by the FTC; some *LB*s correspond to four-digit SIC industries and others correspond to groups of four-digit or to a single three-digit industry. As dependent variables, we use the ratio of company-financed *R&D* expenditures to sales from the 1976 Line of Business data and a measure of innovation derived from responses to two questions in the Levin et al. survey. Respondents were asked to score on a seven-point scale the rates at which new products and new processes were introduced into their *LB*s during the 1970's. We measure innovation in each *LB* by summing the mean scores on these two questions. Concentration ratios are taken from the 1972

Census of Manufacturers, and aggregated to the *LB* level when necessary. Since intertemporal changes in concentration and in *R&D* intensity occur slowly, we test each of our reported specifications for indications of simultaneity.

Consistent with the conventional wisdom, *OLS* and *2SLS* regressions of *R&D* intensity on concentration, its square, and a constant term provide support for the inverted-U hypothesis. Since independence of the concentration variables and the disturbance term is decisively rejected, we report the *2SLS* version (with asymptotic standard errors in parentheses):

$$(1) \quad RD/S = 1.810 \\ (1.402) \\ + 0.166C4 - 0.159(C4SQ \times 100). \\ (0.067) \quad (0.068)$$

The coefficients on *C4* and *C4SQ* are both significant at the .01 level (one-tailed tests). Consistent with the previous literature, in which innovative effort typically reaches a maximum at levels of *C4* between 50 and 60, our results indicate that *R&D* intensity is maximized at a *C4* of 52.

We get similar results by regressing our innovation measure on concentration. Here, however, we report *OLS* estimates, since we cannot reject the hypothesis that concentration and its square are independent of the disturbance term:

$$(2) \quad INNOV = 6.013 \\ (0.790) \\ + 0.089C4 - 0.082(C4SQ \times 100). \\ (0.036) \quad (0.036)$$

The coefficients are again significant at the .01 level, and the rate of innovation reaches a maximum at a *C4* of 54.

Controlling for two-digit industry fixed effects, we find that the coefficients of the concentration terms are somewhat reduced in the *R&D* equation, but both remain significant at the .10 level and *R&D* reaches a maximum at a *C4* of 54. Controlling for fixed effects has virtually no impact on the innovation rate equation. The coefficients on

the concentration terms are essentially unchanged, and they remain significant at the .01 level. The rate of innovation is maximized at a *C4* of 58.

To examine whether the effect of concentration on innovative effort and output is more seriously attenuated when direct measures of technological opportunity and appropriability are included in the regressions, we constructed several variables from the Levin et al. survey data. Analysis of the survey data is still in progress, and undoubtedly the quality of the results reported here can be improved once the statistical properties of the survey responses are more fully understood.

Our variables are intended to capture three dimensions of technological opportunity: closeness to science, external sources of technical knowledge, and industry maturity. To measure closeness to science, we used survey responses concerning the relevance of various basic and applied sciences to the technology of each industry. *SCIENCEBASE* was defined as the maximum score received by any one of the eleven fields of science listed on the questionnaire. We also took account of the contributions to an industry's technical knowledge from four external sources: upstream suppliers of raw materials and equipment (*MATERIALTECH* and *EQUIPTECH*, respectively), downstream users of the industry's products (*USERTECH*), and government agencies and research laboratories (*GOVTECH*). Each of these opportunity variables is measured on a seven-point scale. In light of the widely held view that technological opportunity may vary systematically as an industry matures, we also included a variable intended to capture the relative immaturity of an industry's technology. This was operationalized as the percentage of an industry's property, plant, and equipment that was installed within the five years preceding 1976 (*NEWPLANT*), as reported to the FTC's Line of Business Program.

We used two variables to represent appropriability conditions. The Levin et al. survey asked respondents to characterize (on a seven-point scale) the effectiveness of six mechanisms used by firms to capture and

protect the returns from new processes and new products resulting from *R&D* efforts. The listed mechanisms were patents to prevent duplication, patents to secure royalty income, secrecy, lead time, moving down the learning curve, and complementary sales and service efforts. We limit ourselves here to an attempt to discern whether our sample industries had any effective mechanism of appropriation. We therefore operationalize *APPROPRIABILITY* as the maximum score received by any one of the six mechanisms for either process or product *R&D*.

The survey also asked industries to report on the range of imitation costs and time lags for major and minor, process and product, and patented and unpatented innovations. These measures tend to be highly correlated with one another, though they are not highly correlated with *APPROPRIABILITY*. We use here the average time (in months) required to duplicate a patented, major product innovation (*IMLAG*).

Our principal interest in this paper is the robustness of the results that innovative effort and output depend on concentration. In the absence of an explicit structural model, however, prior expectations about the signs of the opportunity and appropriability variables are in some cases ambiguous. Opportunity, as measured by closeness to science and the strength of other external influences on technology, should have a positive effect on innovative output, but to the extent that the efforts of upstream suppliers, downstream users, and the government substitute for an industry's own *R&D* effort, these variables may have a negative sign in the *R&D* equation. Appropriability may also have ambiguous effects on *R&D* incentives, as recent theoretical work has emphasized. Some effective means of appropriating returns is surely required to elicit *R&D* effort. Spillovers among competitors, however, are ambiguous in their effects. They create a disincentive to *R&D* investment by dissipating the innovator's rents, but they also enhance the productivity of *R&D* by strengthening the industry's knowledge base (see Michael Spence, 1984). Thus, *APPROPRIABILITY*, which measures the extent to which an industry has at least one effective means of appropriation,



TABLE 1—DETERMINANTS OF R&D INTENSITY  
(2SLS ESTIMATES)

	Regression Coefficients	
	(1)	(2)
Intercept	-4.650 <sup>c</sup> (2.324)	<sup>a</sup>
C4	0.022 (0.082)	-0.005 (0.073)
C4SQ × 100	-0.013 (0.082)	0.002 (0.072)
NEWPLANT	0.048 <sup>b</sup> (0.015)	0.025 <sup>b</sup> (0.014)
SCIENCEBASE	0.469 <sup>b</sup> (0.215)	0.490 <sup>b</sup> (0.202)
MATERIALTECH	-0.043 (0.163)	-0.037 (0.154)
EQUIPTECH	-0.368 (0.190)	0.164 (0.185)
USERTECH	0.365 <sup>c</sup> (0.150)	0.029 (0.152)
GOVTECH	0.304 <sup>c</sup> (0.129)	0.282 <sup>c</sup> (0.139)
APPROPRIABILITY	0.069 (0.258)	0.196 (0.243)
IMLAG	0.052 (0.073)	0.100 (0.068)
Chi-Square Tests (D.F.):		
All survey vars. (7)	3.713 <sup>d</sup>	2.487 <sup>d</sup>
Opportunity vars. (5)	4.820 <sup>d</sup>	2.917 <sup>d</sup>
Appropri. vars. (2)	0.287	1.395
Industry dummies (13)		3.864 <sup>d</sup>

Note: Asymptotic standard errors are shown in parentheses.

<sup>a</sup>Separate intercepts for 14 industry groups.

<sup>b</sup>Significant at .05 level (one-tailed *t*-test).

<sup>c</sup>Significant at .05 level (two-tailed *t*-test).

<sup>d</sup>Significant at .05 level (*Chi*-square test).

TABLE 2—DETERMINANTS OF THE RATE OF INNOVATION  
(OLS ESTIMATES)

	Regression Coefficients	
	(1)	(2)
Intercept	-4.893 (2.558)	<sup>a</sup>
C4	0.055 (0.037)	0.048 (0.038)
C4SQ × 100	-0.039 (0.038)	-0.036 (0.038)
NEWPLANT	0.042 <sup>b</sup> (0.018)	0.017 (0.019)
SCIENCEBASE	0.371 (0.240)	0.439 <sup>b</sup> (0.253)
MATERIALTECH	0.388 <sup>b</sup> (0.184)	0.294 (0.191)
EQUIPTECH	0.246 (0.204)	0.472 <sup>b</sup> (0.234)
USERTECH	0.277 (0.172)	0.035 (0.195)
GOVTECH	0.048 (0.148)	0.051 (0.178)
APPROPRIABILITY	0.564 <sup>b</sup> (0.266)	0.745 <sup>b</sup> (0.279)
IMLAG	-0.035 (0.082)	-0.025 (0.085)
F-Tests (D.F.): <sup>c</sup>		
All survey vars. (7)	2.797 <sup>d</sup>	2.682 <sup>d</sup>
Opportunity vars. (5)	3.288 <sup>d</sup>	2.633 <sup>d</sup>
Appropri. vars. (2)	2.307	3.573 <sup>d</sup>
Industry dummies (13)		1.640

Note: Standard errors are shown in parentheses.

<sup>a</sup>Separate intercepts for 14 industry groups.

<sup>b</sup>Significant at .05 level (one-tailed *t*-test).

<sup>c</sup>Numerator D.F. in parentheses; denominator D.F. is 116 for col. (1) and 103 for col. (2).

<sup>d</sup>Significant at .05 level (*F*-test).

should have a positive effect on both innovative effort and output. On the other hand, a long imitation lag, which inhibits spillovers, may encourage innovative effort by reducing the disincentive effect, but it may also reduce innovative output, by reducing the productivity of R&D.

Table 1 displays estimates of the R&D intensity equation, alternately excluding and including two-digit industry fixed effects. We again report 2SLS results, because the hypothesis of independence between the concentration terms and the disturbance is rejected by Wu tests. We also found no evidence of heteroskedasticity, despite the

fact that the number of responses to the Levin et al. survey varied across industries.

Comparing Table 1 with (1) in the above text, the coefficients and *t*-statistics on the concentration terms fall by an order of magnitude. There are no anomalous signs; the survey variables as a whole and the opportunity measures in particular are highly significant. Specifically, R&D spending appears to be encouraged in youthful industries where a strong science base is present and where the government makes substantial contributions to technological knowledge. The continued statistical significance of these variables when fixed industry effects are taken

into account suggests that we may have isolated some relevant dimensions of technological opportunity.

Table 2 presents *OLS* estimates of the innovation equation. The *2SLS* estimates were substantively identical, and we found no evidence of simultaneity. Once again, the concentration terms become statistically insignificant at the .05 level once the survey variables are added to the equation, although the linear term is significant at the .10 level. All survey variables have the predicted sign, and they are jointly significant at the .05 level. The two-digit industry effects are jointly insignificant, but their inclusion strengthens the performance of several survey variables: *SCIENCEBASE*, *EQUIPTECH*, and *APPROPRIABILITY*. The signs of the coefficients on *APPROPRIABILITY* and *IMLAG* conform to the expectations discussed above.

### III. Conclusions

In the spirit of "creative destruction," we hope that our findings will at last move the empirical literature beyond the oversimplified propositions that industrial concentration promotes innovative effort and innovative output. To explain interindustry variation in *R&D* incentives and the productivity of innovative effort, we must look to underlying differences in technological opportunities and appropriability conditions. In this paper, we have made only illustrative

use of the Levin et al. data. These new data, and a burgeoning theoretical literature on *R&D* competition in the presence of imperfect appropriability, offer encouragement that a deeper understanding of industrial innovation is within our grasp.

### REFERENCES

- Levin, Richard C., Klevorick, Alvin K., Nelson, Richard R. and Winter, Sidney G., "Survey Research on R&D Appropriability and Technological Opportunity: Part 1," Working Paper, Yale University, July 1984.
- \_\_\_\_\_ and Reiss, Peter C., Tests of a Schumpeterian Model of R&D and Market Structure," in Zvi Griliches, ed., *R&D, Patents, and Productivity*, Chicago: University of Chicago Press, 1984, 175-204.
- Scherer, F. M., "Market Structure and the Employment of Scientists and Engineers," *American Economic Review*, June 1967, 57, 524-31.
- Schumpeter, Joseph A., *Capitalism, Socialism, and Democracy*, 3rd ed., New York: Harper & Row, 1950.
- Scott, John T., "Firm versus Industry Variability in R&D Intensity," in Zvi Griliches, ed., *R&D, Patents, and Productivity*, Chicago: University of Chicago Press, 1984, 233-48.
- Spence, Michael, "Cost Reduction, Competition, and Industry Performance," *Econometrica*, January 1984, 52, 101-21.

# Patent Licensing and R&D Rivalry

By CARL SHAPIRO\*

It is widely recognized that long-run industrial performance is influenced not only by static efficiency, but also by the rate of technological progress. The single most important factor determining the rate of technological progress in a given industry is the level of industrywide expenditures on research and development (*R&D*) activities.

Another critical factor in determining the rate of technological progress is the rate of *diffusion* of new technologies. The pattern of diffusion influences the pace of technological progress in two ways. First, it directly affects the costs of noninnovating firms and hence the speed with which technological advances are utilized throughout an industry. Second, it feeds back onto the incentives of firms to engage in *R&D*. It is useful to distinguish these two types of effects; I will call the first the *ex post* effect of diffusion and the second the *ex ante* effect.

There are three basic channels through which dissemination takes place: patent licensing, research joint ventures, and imitation. Licensing is a voluntary form of dissemination whereby an inventor can enjoy at least some of the gains to trade from spreading the use of his superior technology. Research joint ventures (*RJVs*) are a way of agreeing to share research results *before* they are realized, a type of "*ex ante* licensing." Imitation is a form of diffusion over which the patentee has little control; noninnovating firms unilaterally appropriate some of the benefits of the discovery.

Intuition suggests that licensing has socially beneficial *ex post* effects, and that it

encourages innovation (*ex ante*) by increasing the rewards to the patentee. Research joint ventures also appear to have socially, if not privately, beneficial effects upon *ex post* dissemination, but it is not clear offhand what effect *RJVs* have on development incentives. Finally, one expects that imitation will have socially favorable *ex post* effects, but will tend to discourage or retard innovation.

In this paper I indicate how the intuition expressed in the previous paragraph stands up under closer analysis. I will discuss the factors that appear most important in determining both the pattern of dissemination of innovations, and the social and private effects of such dissemination. I will emphasize the effects of patent licensing, touching upon joint ventures and imitation along the way. The analysis will be conducted in the context of an oligopolistic industry. An oligopolistic setting is the natural one on which to focus, as patents, by their very nature, create market power, and as many progressive industries have concentrated market structures.

## I. Licensing of a Given Innovation (*Ex Post* Licensing)

Once an innovation has been patented, there are natural social and private incentives to see that it is employed widely in situations where it will lower production costs. With perfect information, and no transactions costs, therefore, we would expect licenses to be issued to all firms that could benefit from the improved technology.

The argument is straightforward: there are gains to trade from licensing, and the patentee can design contracts that split these gains between itself and potential licensees. For example, it could charge a royalty per unit of output that just equals the reduction in per unit costs that the licensee achieves on account of the new technology. This would leave the licensee indifferent between licens-

\*Princeton University, Princeton, NJ 08540. The reader will soon observe my debt to my colleague, Michael Katz, with whom I have done much of the work on patent licensing described herein. I thank Nancy Gallini for extensive comments on related work, Richard Caves for informing me about patent licensing in practice, Paul David, Avinash Dixit, Gene Grossman, and Ken Judd for comments and suggestions on a draft of this paper, and Steve Salop for general inspiration.

ing and not, and let the licensor enjoy the gains from the licensee's improved efficiency. While the licensor appropriates all of the gains to trade with this particular contract, there are clearly other contracts that split the gains more equally.

What will the (privately) *optimal* licensing contract look like? In general, patent licensing contracts can be designed to implement perfectly the monopoly (fully collusive) outcome. Doing so will maximize the pie that is to be split between the licensor and any licensees. To achieve the collusive outcome, it is sufficient to consider two-part tariff contracts consisting of a fixed fee,  $F$ , and a per unit royalty rate,  $r$ .

This general principle is most transparent in the context of a single firm licensing its patent to its lone rival. The idea is that the licensing firm can control its rival's marginal cost, and hence can manipulate its rival's behavior (reaction curve), through the use of the per unit royalty rate. With an appropriate choice of royalty rate, the collusive outcome can be achieved. Doing so is simply a matter of computing the licensee's reaction curve, given the royalty rate, and then choosing that rate so that the licensee's reaction curve intersects that of the licensor at a point where industry output is at the monopoly level.

Two rivals (with or without innovations) alternatively could design a cross-licensing agreement whereby each would pay the other a royalty per unit of output, ostensibly for the right to use the other's technology. By imposing a "tax" on each other (or by writing pricing or territorial restrictions into the cross-licensing contract), the firms could again achieve the fully collusive outcome. A cross-licensing contract may be required to achieve the fully collusive outcome if the firms produce different products or are otherwise heterogeneous.

The reader may be surprised that even a minor innovation can markedly alter the industry's final behavior. Indeed, a sham innovation can be "licensed" to facilitate collusion! How is this possible? For such an "innovation," the licensing contract would have a large royalty rate and a negative fixed fee. That is, the licensor would reduce its rival's output by imposing a "tax" of  $r$  per

unit, and then compensate the licensee for this tax with a negative fixed fee (i.e., a side payment). In the extreme case, this licensing contract would be equivalent to a bribe paid by the licensor to induce the licensee to exit the industry.

Such a side payment, in exchange for which the licensee would reduce its output, is likely to be illegal under the antitrust laws, and for a good reason! So, a reasonable constraint to put on the two-part tariff contract is that the fixed fee be nonnegative. Equivalently, the licensor cannot charge a royalty rate in excess of the per unit savings the licensee enjoys with the new technology; the licensing contract cannot raise the licensee's marginal (production plus royalty) costs. With this restriction, the optimal contract would set the royalty rate as high as possible (i.e., at the per unit cost savings afforded by the new technology), and have no fixed fee.<sup>1</sup>

My analysis so far indicates that licensing contracts should arise for any innovation for which the transactions costs are low, and two-part tariff contracts can be written. Yet licensing among rivals is not often observed in practice. Besides the direct costs of technology transfer, there are three information problems that may limit the firms' ability to achieve the licensing gains to trade. First, there may be asymmetric information regarding the value of the license; this may undermine the parties' ability to strike a licensing deal. Second, the innovator may find it difficult to let others use his invention without giving them useful information in the ongoing competition to acquire additional patents.<sup>2</sup> Third, it may be costly or

<sup>1</sup> Nancy Gallini and Ralph Winter (1984) and Morton Kamien and Yair Tauman (1984) analyze licensing contracts of this form. Such a contract gives all of the licensing gains to trade to the licensor. Depending upon the bargaining power of the two parties, a lower royalty may arise. According to Richard Caves et al. (1983), licensors enjoy an average of about 40 percent of the gains to trade from licensing.

<sup>2</sup> It may be difficult to design a contract that prevents such usage of the information. In a survey reported by Caves et al., 43 percent of licensing agreements contained a "technology flowback" provision, which required the licensee to share any advances or improvements in the licensed technology with the licensor. For

impossible for the licensor to monitor the licensee's output so as to charge per unit royalties. The first two considerations are beyond the scope of this paper; I shall, however, report some results relating to the third.

If per unit royalties cannot be levied, the licensor must balance off two effects of a fixed fee license: such a license spreads the use of the improved technology, but it also makes the licensee a more vigorous rival. We would expect a license to be issued if and only if doing so raises the joint profits of the licensor and the licensee (in which case they could split the gains in some fashion). In the duopoly context, joint profits are industry profits, so the question of whether licensing will occur reduces to the question of how lowering the licensee's costs influences industry profits.

In this context, Michael Katz and my paper (1984a) shows that drastic innovations (those allowing the patentee to enjoy a monopoly) will not be licensed, and that an innovation that is nearly drastic will also not be licensed. On the other hand, a small innovation by one of two equally efficient firms will be licensed if and only if  $x p''(x)/p'(x) > -2$ , where  $p(x)$  is the industry demand curve, and  $x$  is the initial industry output level. For linear demand ( $p'' = 0$ ), therefore, small innovations will be licensed. This suggests that the neglect of patent licensing in the existing R&D literature is unfounded: even under stringent constraints on the licensing contract, licensing is often profitable. This result would only be strengthened with more than two firms, as the other firms would reduce their output in response to such a licensing agreement.

When licensing cannot raise the licensee's marginal costs, as is the case for fixed fee licensing, it will tend to increase the licensee's output, and total industry output as well. So, consumers will benefit from patent licensing. In the duopoly context, this implies that the firms' private incentives to license are less than the social incentives. In an oligopoly

context this conclusion is not warranted, however, as the firms that are not parties to the licensing contract will typically be made worse off on account of the license. A final welfare point is that (fixed fee) licensing is not always socially beneficial. (See Katz and myself, 1984a.) If the licensor is more efficient than the licensee, even with the new technology, it may be optimal for the licensee to produce using the old technology. This will definitely be the case if the licensee would be close to shutting down even with the license. The reason is that licensing to the less efficient firm diverts output from the more efficient firm under a variety of oligopoly solution concepts. So, even *ex post*, and even disallowing per unit royalties, mandatory licensing may be undesirable.

So far I have discussed mainly one duopolist licensing its innovation to its rival. A new set of issues arises when one considers the optimal way for an innovator to license his patent to a subset of (downstream) oligopolists. The added richness arises because one firm's willingness to pay for a license depends upon the set of other firms that are also purchasing licenses. This *interdependence of demands* is a consequence of the fact that the new technology reduces each licensee's marginal cost, and hence alters its reaction curve in the final pricing or output game.

In general, the value to a given oligopolist of obtaining a license to a product or process innovation depends upon the set of the firm's rivals purchasing licenses. For process innovations in a homogeneous good industry, obtaining a (fixed fee) license is *less* valuable to a given firm if more of its rivals have done so. If the license permits the firm to adopt a new industry standard, however, it may be *more* valuable if more rivals have licenses. Both the optimal licensing scheme and the pattern of dissemination to which it leads depend heavily upon whether the demand interdependencies are "negative" or "positive."

Morton Kamien and Yair Tauman have analyzed licensing in the context of a symmetric, Cournot oligopoly with constant average costs facing a linear demand schedule. They consider fixed fee contracts and pure per unit royalty contracts. The optimal

---

further information on patent licensing contracts in practice, including the use of fixed fees vs. per unit royalties, see Charles Taylor and Z. A. Silberman (1973).

royalty contract charges a royalty that just equals the per unit cost savings from the new technology (as noted above in a duopoly context). With this contract, each licensee's downstream behavior is unaffected by the contract, and the demand interdependencies vanish. Kamien and Tauman demonstrate that fixed fee contracts are more profitable for the patentee than are royalty contracts. The reason is that they *do* exploit the demand interdependencies, leaving the firms *worse off* than they were prior to the innovation.

Restricting attention to fixed fee licensing, Katz and I (1984b) have extended this analysis in a number of directions. We consider very general downstream oligopoly cost, demand, and behavioral conditions, and general auction schemes for selling the licenses. Furthermore, we look both at the case where the patent is owned by an independent (upstream) research lab, and the case where the patent is jointly owned by a subset of the (downstream) oligopolists, a research joint venture. One principle that emerges is that it is generally *not* optimal to sell the licenses using the conventional price system. That is, it is not optimal (for the patentee) simply to set a price per license and let the firms each decide whether or not to purchase a license at the set price.

The reason the price system does not maximize the seller's revenues is that it has no built-in ability to "threaten" the buyers. When one buyer decides not to purchase a license, his decision has no effect on the purchase pattern of the other firms. So, under the price system, a buyer can take as given the number of others that will purchase licenses. If, in equilibrium,  $k$  of the  $n$  oligopolists purchase licenses, the most that each buyer will pay for a license is  $W(k) - L(k - 1)$ , where  $W(k)$  is the payoff to a license holder (a "winner"), gross of any license fees, if  $k$  licenses are sold, and  $L(k)$  is the payoff to a firm without a license (a "loser") if  $k$  are sold.

Consider the alternative "quantity" system, whereby the seller announces that  $k$  licenses are to be sold to the  $k$  highest bidders. Then, a firm knows that  $k$  licenses are sure to be sold, and it will pay up to

$W(k) - L(k)$  for one. So long as a non-licensee is worse off the more of its rivals purchase licenses (i.e., so long as  $L(k)$  is strictly decreasing with  $k$ ), the quantity system will yield the seller higher revenues.

Katz and I (1984b) characterize the inventor's optimal licensing strategy within the following class of sales mechanisms: the inventor can run an auction in which there are  $k$  licenses available, and these will be sold to the  $k$  highest bidders (at their bidding prices), subject to a minimum bid. Making licenses available for all potential buyers at a minimum bid  $b$  amounts to using the price system with a price of  $b$  per license.

When the inventor employs the optimal auction scheme to sell the licenses, all of the firms that are not members of the joint venture (which is *all* of the oligopolists in the case of the independent research lab) are made worse off on account of the innovation. In terms of the notation introduced above, the oligopolists' pre-innovation payoff levels are given by  $L(0)$ ; after the innovation, the firms earn the lower payoff of  $L(k)$  if  $k < n$  licenses are sold.<sup>3</sup> Essentially, the innovator appropriates some of the oligopolists' initial property rights to their oligopoly profits, by threatening them with the lower profit level that they would earn if some of their rivals obtained licenses and they did not.

Katz and I (1984b) also show how the pattern of ownership of the patent influences its dissemination. In the class of mechanisms described above, larger joint ventures have an incentive to issue fewer licenses. So, the notion that *RJVs* promote dissemination may be unfounded. Even if the joint venture is contractually required to issue licenses to all of its members, the independent lab will issue more licenses than will a small joint venture.<sup>4</sup>

<sup>3</sup>If it is optimal to issue licenses to all  $n$  of the oligopolists, then the inventor *will* employ the price system, and each firm's payoff will be  $L(n - 1)$ .

<sup>4</sup>Of course, large joint ventures will disseminate the innovation more widely if they must give it to all of their members on a royalty free basis. Also, if entry fees can be charged in the license auction, ventures of all sizes will have incentives to issue the number of licenses that maximizes industrywide oligopoly profits.

If one permits the patent holder to issue two-part tariff licensing contracts, and to use a general auction mechanism with entry fees, the industry will duplicate the fully collusive outcome under the new technology. The patent holder will appropriate all of these profits for himself, and the oligopolists (not in the venture) will earn zero profits, after licensing royalties are accounted for. So licensing, while spreading the use of the new technology, may lead to less competitive industry behavior.

## II. The Effect of Licensing on Development Incentives (*Ex Ante* Effects)

If one considers a single (potential) inventor investing resources to develop a given innovation, it is clear that the possibility of licensing can only increase the value that the inventor places upon developing the invention. After all, he can always refrain from licensing, and often will find the payoff to innovating higher on account of licensing.

In addition to increasing the (single) inventor's development incentives, licensing may make these incentives socially excessive. There are two reasons why this may happen. First, there are private returns to patenting, namely the induced monopolization of the downstream industry through the use of per unit royalties, which are not social returns. This reward for the inventor may be viewed as a return to the ability to write contracts among the oligopolists that raise their perceived marginal costs. Second, even without the use of such royalties, the private returns to the inventor include the reduced profits of the downstream oligopolists, as noted above. These are transfers, not social benefits arising from the innovation.

When competition to develop an innovation is considered, the effect of licensing upon R&D is not necessarily positive. First, in the context of multiple innovations, licensing may be used to strategically deter a rival's R&D. Nancy Gallini (1984), and Gallini and Ralph Winter, show that such strategic (*ex ante*) licensing is more likely to occur when the firms' costs differ widely.

Even *ex post* licensing need not encourage R&D in a patent race context. While it is

true that licensing increases the returns to winning the patent, it also increases the returns to *not* winning the patent. Specifically, a firm may know that it will become a licensee if it loses the patent race. So long as the licensor does not appropriate all of the gains to trade from licensing, the expectation of being a licensee will dampen a firm's incentives to win the patent. Clearly, the way in which the gains to trade from licensing are split between the licensor and the licensee is a critical factor influencing firms' incentives to develop.

In a duopoly patent race model, Katz and I (1984c) show that the possibility of (*ex post*) licensing can completely alter the nature of dynamic rivalry. Specifically, each firm may prefer losing the race to winning it. Likewise, each firm may *benefit* when its rival patents. The former will occur if the development and patenting costs (which are incurred only by the patentee) are large, and the latter if the licensee enjoys substantial benefits on account of the licensing contract. In such cases, the race takes on the character of a waiting game: each firm hopes that its rival will develop the innovation, which is like a public good for the two firms. (The same effect can also arise due to imitation rather than licensing.) Licensing may delay innovation, as each firm is more content to wait, lose the race, and become a licensee.

## III. Conclusion

Patent licensing, research joint ventures, and imitation have significant effects on both the rate of (*ex post*) diffusion of new technologies, and on private firms' (*ex ante*) incentives to develop such technologies. Recent work analyzing the impact of these channels of dissemination indicates that these *ex ante* and *ex post* effects interact in surprising ways. For example, while licensing contracts serve a useful function of spreading the use of superior technologies, they may be used to facilitate collusion. While patent licensing increases the value of winning a patent, it also increases the value of losing a patent race, and thus need not encourage innovation. And, while research joint ventures avoid duplication of research efforts

(see Katz, 1984, or Gene Grossman and myself, 1984), they may inhibit diffusion and reduce development incentives (Katz and myself, 1984b).

While much progress has been made in understanding dissemination and its role in the process of technological change, large gaps remain in our understanding. Most notably, more work is needed to understand the information problems facing patent licensing contracts, and to understand the effect of research joint ventures on the pace of technological advance.

#### REFERENCES

- Caves, Richard, Crookell, Harold and Killing, J. Peter, "The Imperfect Market for Technology Licenses," *Oxford Bulletin of Economics and Statistics*, August 1983, 45, 249-67.
- Gallini, Nancy "Deterrence by Market Sharing: A Strategic Incentive for Licensing," *American Economic Review*, December 1984, 74, 931-41.
- \_\_\_\_\_ and Winter, Ralph, "Licensing in the Theory of Innovation," University of Toronto, April 1984.
- Grossman, Gene and Shapiro, Carl, "Research Joint Ventures: An Antitrust Analysis," Woodrow Wilson School Working Paper No. 68, Princeton University, January 1984.
- Kamien, Morton and Tauman, Yair, "Fees vs. Royalties and the Private Value of a Patent," Northwestern University, September, 1984.
- Katz, Michael, "An Analysis of Cooperative Research and Development," Woodrow Wilson School Working Paper No. 76, Princeton University, May 1984.
- \_\_\_\_\_ and Shapiro, Carl, (1984a) "On the Licensing of Innovations," Woodrow Wilson School Working Paper No. 82, Princeton University, October 1984.
- \_\_\_\_\_ and \_\_\_\_\_, (1984b) "How to License a Patent," Woodrow Wilson School Working Paper No. 84, Princeton University, December 1984.
- \_\_\_\_\_ and \_\_\_\_\_, (1984c) "Perfect Equilibrium in a Development Game with Licensing or Imitation," Woodrow Wilson School Working Paper No. 85, Princeton University, December 1984.
- Taylor, Charles T. and Silberston, Z. A., *The Economic Impact of the Patent System*, Cambridge: Cambridge University Press, 1973.



## Nominal Wage-Price Rigidity as a Rational Expectations Equilibrium

By COSTAS AZARIADIS AND RUSSELL COOPER\*

Most recent studies of macroeconomic behavior fall into one of two categories. The first, often called the equilibrium business cycle approach, stems from the fundamental contribution of Robert E. Lucas (1972), and related work by Thomas Sargent and Neil Wallace (1975) and many others. These studies espouse the view that a positive correlation between output and the stock of paper assets can arise if households are unable to identify the source and, therefore, the permanence of price movements.

Employment and output responses in this view are driven by the intertemporal substitution effect, especially the substitution of current leisure for future consumption. Since cyclical fluctuations in employment are large relative to the corresponding real wage movements, substantial wage elasticity of labor supply is required to validate these models.

The equilibrium approach to business cycles implies certain restrictions on the conduct of monetary policy. In particular, rational expectations undermine the ability of the monetary authority to influence economic activity in a systematic manner; see Sargent-Wallace for an example.

Sticky wages and prices are the cornerstone of an alternative description of macroeconomic behavior. Rooted vaguely in Keynes, and more firmly in the "dual decision hypothesis" of Robert Clower (1984), this approach studies equilibria with quan-

tity rationing; see Edmond Malinvaud (1977). The rationing story lacks a precise specification of the *source* of price stickiness, offering very little guidance about the eventual causes of price change.

Considerable efforts were made in the 1970's to fill this lacuna in Keynesian macroeconomics. Beginning with work by Martin N. Baily (1974) and others, the implicit contracts literature focused on the incomplete insurability of human capital. Unable to find insurance against fluctuations in labor income elsewhere, workers demand insurance from those best placed to observe labor income—their own employers. Both wage inflexibility and layoffs, then, can be viewed as an outcome of a joint insurance-employment relationship between workers and firms; see Azariadis (1975).

Critics like George Akerlof and Hajime Miyazaki (1980) soon discovered that the original contracting models could not produce layoffs without prohibiting severance pay or otherwise limiting the terms of the contract. Others pointed out that these models were determinedly microeconomic, offering few insights into the stickiness of *nominal* wages or the effectiveness of stabilization policy.

Two quite distinct lines of research developed out of the original implicit contract ideas. One focuses on asymmetric information and implementability (see the *QJE* 1983 Symposium for original work and the review article by Oliver Hart, 1983) as a means of driving a wedge between the *ex post* marginal rates of substitution of the contractants. Under some technical assumptions, the outcome is "involuntary" underemployment or unemployment.

We are most concerned here with the other line of research, which sought to fit labor or

<sup>†</sup>*Discussants:* Guillermo Calvo, Columbia University; Jo Anna Gray, Washington State University.

\*Department of Economics, University of Pennsylvania, Philadelphia, PA 19104; and Cowles Foundation, Yale University, New Haven, CT 06520, respectively. We thank Takatoshi Ito for useful comments.

commodity contracts into full-blown macroeconomic structures; see Stanley Fischer (1977), and Arthur Okun (1981). These authors apparently took their cue from the insurance arrangements stressed in the implicit contracts literature. Just as in the rationing approach mentioned above, the macro contracts literature does not fully derive the structure of trades from first principles; it asserts instead that money wages are predetermined and that employment is set by labor demand. This rule (see Robert Barro, 1977) is not in the best interest of the contractants for it requires laborers to be permanently off their offer curve. Consequently, macroeconomic contract models are not an entirely reliable guide for policy evaluation.

As we take stock of matters, it seems to us that the time is ripe to assess what role contracts play in macroeconomics. An ideal outcome would be a coherent story of money-wage stickiness, and of output response to stabilization policy.

This paper distills recent research, by ourselves and others, that pursues that ideal without quite reaching it. The presentation here favors intuition over technical detail as much as possible. The next section, in particular, discusses examples of intertemporal economies that possess competitive rational expectations equilibria in which wages and prices are predetermined, output is sensitive to policy shocks, and markets clear. Section II reports on the existence of Nash equilibria with similar properties when producers set both prices and wages.

# I

We outline here some rational expectations equilibria of an overlapping generations economy that begins at time zero and goes on forever. The beginning of each time period  $t = 0, 1, \dots$ , coincides with the appearance of a generation, labeled "generation  $t$ ," that consists of  $2N$  individuals, all of whom survive for two full periods: "youth" and "old age." Individuals specialize over their life cycle, participating in the production of a single, perishable commodity when young, saving their entire factor rewards to purchase a paper asset called "money," and

using their assets in old age to finance the consumption of the same commodity they helped produce in youth. We assume that no endowments exist of this valuable commodity—it must be produced before it is consumed.

Every generation consists of  $N$  identical, risk-neutral workers, and  $N$  identical, risk-neutral entrepreneurs. Each worker is endowed with  $\bar{n} > 0$  units of divisible leisure in youth, and with a utility function  $c_{t+1} - kn_t$ , over old age consumption ( $c_{t+1}$ ) and youthful work ( $n_t$ ). The parameter  $k$ , interpreted here as a reservation wage, is the same for all workers in all generations, and lies in the interval  $(0, 1)$ .

Entrepreneurs own in youth the constant-returns technology  $y_t = n_t$ , the only means of production in this society. Denoting by  $p_t$  the money price of goods and by  $w_t$  the money wage, it is clear that a generation  $t$  entrepreneur consumes in old age  $(p_t - w_t)n_t/p_{t+1}$ .

Money is printed by the government to finance purchases of goods from private sellers. Real per capita government purchases are an independent, identically distributed random variable,  $\tilde{g}_t$ , with a probability measure defined over the positive interval  $[a, b]$ , and expected value  $\mu > 0$ . We assume  $b < \bar{n}$ , that is, government purchases never exceed the productive capacity of the economy.

If  $M_t$  is the nominal stock of money in period  $t$  and  $m_t = M_t/p_t$  is its purchasing power, the government budget constraint can be written in either of the following two forms:

$$(1) \quad \begin{aligned} p_t g_t &= M_t - M_{t-1} \\ p_{t-1}/p_t &= (m_t - g_t)/m_{t-1}. \end{aligned}$$

All agents, including the government, are wage and price takers. Commodities purchased by the government are used up with no effect on any household's preferences over private goods.

The constant returns to scale assumption means that the demand for labor (and supply of goods) is zero if  $p_t < w_t$ , and infinity if  $p_t > w_t$ . Competitive equilibrium is trivial in the former case, and nonexistent in the latter.

We focus therefore on the case  $p_t = w_t$ , which makes employment supply determined. In particular,  $n_t \geq 0$  if the expected purchasing power of wages satisfies

$$(2) \quad w_t E_t(1/p_{t+1}) = p_t E_t(1/p_{t+1}) \geq k.$$

Also,  $n_t = \bar{n}$  if (2) holds as a strict inequality. Here  $E_t$  is the expectations operator conditional on all events up to, and including, period  $t$ .

Equilibrium in the goods market requires goods supply  $n_t$  to equal goods demand. Demand is simply the real value of existing money balances  $m_t$ , held by retired individuals and by the government.

Combining the market-clearing condition  $m_t = n_t$  with relations (1) and (2), we see readily that every competitive equilibrium satisfies

$$(3) \quad E_t m_{t+1} \geq k m_t + \mu,$$

with  $m_t \leq \bar{n}$ , and  $m_t = \bar{n}$  if (3) is a strict inequality. As one may guess, inequality (3) admits many equilibria, the most interesting of which we present below (see also Roger Farmer and M. Woodford, 1984).

One class of equilibria worth noting features *flexible prices and predetermined quantities*: current events are reflected in current equilibrium prices but not in current equilibrium quantities. A member of this class is the full-employment equilibrium ( $p_t = w_t$ ,  $p_{t-1}/p_t = 1 - g_t/\bar{n}$ ,  $m_t = \bar{n}$ ) for all  $t$ , which is easily seen to satisfy relations (1) and (3) provided that

$$(4) \quad \mu \leq (1 - k)\bar{n}.$$

A second class of equilibria has *predetermined prices and flexible quantities*: current events influence current quantities but not prices. One member of this class (and the only one existing in our simple economy) is ( $p_t = w_t$ ,  $p_{t-1}/p_t = k$ ,  $m_t = k m_{t-1} + g_t$ ) for all  $t$ . This one, too, is consistent with relations (1) and (3). The implied path of output

$$(5) \quad m_t = k^t m_0 + \sum_{s=0}^{t-1} k^s g_{t-s},$$

stays in the interval  $[a/(1-k), b/(1-k)]$  if it begins there, that is, if the initial parameter  $m_0$  lies in that interval. Therefore, a predetermined wage-price equilibrium exists if

$$(6) \quad b \leq (1 - k)\bar{n}; \quad a > b(1 - k).$$

That is, if the aggregate demand for goods never exceeds the economy's productive capacity, and output never falls short of government consumption.

Both the full-employment and the underemployment equilibria are full rational expectations equilibria in cleared markets. But they differ substantially in what they predict about the economic consequences of policy shocks. The full-employment equilibrium has very classical properties: prices are flexible and complete crowding-out obtains. Government expenditure ( $g_t$ ) displaces private consumption ( $\bar{n} - g_t$ ) one-for-one.

The underemployment equilibrium possesses Keynesian features: predetermined wages and prices; no crowding-out in the short run, since private consumption is predetermined; and a long-run "multiplier" of  $1/(1-k)$ , which exceeds unity. To understand the last assertion, replace the random variable  $\tilde{g}$  by  $\lambda \tilde{g}$ , where  $\lambda$  is a positive constant. Then, clearly, the output path will stay in the interval  $[\lambda a/(1-k), \lambda b/(1-k)]$  if it begins there.

Looking back at inequalities (4) and (6), one sees that neither equilibrium exists if government purchases are "too large" on average; in particular, a predetermined price-and-wage equilibrium is impossible if government purchases are "too dispersed." Wage rigidity is not a frequently observed phenomenon when macroeconomic policy is highly uncertain.

Full-employment equilibria here yield the highest possible consumption, net of the disutility of work, and therefore are superior to underemployment equilibria in the familiar sense of first-degree stochastic dominance. Nevertheless, full employment is but one of many possible equilibria in this model economy. The underdeterminate nature of equilibrium is a weakness of our model, and perhaps of actual intertemporal economies as well. To pin down the equilibrium, we need

an additional restriction, for example, a description of how a hypothetically independent monetary authority would react to government spending. Suppose, in particular, that the authority follows the rule

$$(7) \quad M_t - M_{t-1} = (p_{t-1}/k)g_t.$$

Then predetermined prices are validated and a predetermined quantity equilibrium is no longer tenable.

## II

The above analysis suggests that contracts are not necessary for the existence of equilibria with predetermined wages and prices. In both the predetermined-price underemployment equilibrium and the flexible-price full-employment equilibrium, workers' old age consumption was independent of shocks in their youth. Hence, even if worker's preferences were strictly concave functions of  $c_{t+1} - kn_t$ , labor contracts could not improve the sharing of consumption risks. There are, of course, other spot equilibria in the model of the previous section for which old age consumption does depend on shocks in youth; introducing contracts will generally affect these equilibria in a way that need not concern us here.

A more interesting question is whether the old age consumption of risk-averse workers can be set independently of the government policy shocks realized in their *old age*. Since labor contracts can only protect workers against events in their youth, we consider both commodity and labor contracts to explore fuller lifetime insurance opportunities. This represents a merger of labor contracts with the "customer market" approach advocated by Okun. Using a variant of the model outlined above, we explore whether nominal wage rigidities may be the outcome of a game between wage-setting and price-setting agents.

Assume workers have lifetime utility functions of the form  $u(c_{t+1} - kn_t)$  where  $u$  is increasing and strictly concave. Entrepreneurs of generation  $t$  are still risk neutral, but they are assumed to be born at the beginning of period  $t - 1$  and live for three

full periods, producing in the middle one and consuming in the last one. They are endowed with  $\bar{e} > 0$  of commodities (or leisure) in middle age;  $\bar{e}$  is also the fixed cost of operating a firm.

The source of uncertainty here is slightly different from our first model. The period  $t$  money supply is  $m_t = m_{t-1}\tilde{x}_t$  where  $\tilde{x}_t$  is a random variable with a stationary distribution over the interval  $[1, \bar{x}]$ . All money creation finances government purchases.

We assume that all *ex post* trades are mediated through contracts. A generation  $t$  entrepreneur is involved in *three* contracts. First, he supplies commodities to both old (generation  $t - 1$ ) agents and the government in period  $t$ , under a price schedule established in period  $t - 1$ , before  $x_{t-1}$  is realized. Second, he signs a labor contract with generation  $t$  workers to produce the commodities demanded by the previous generation given the firm's price schedule. Finally, as a future consumer, the entrepreneur makes purchases from a generation  $t + 1$  firm. A generation  $t$  worker faces a period  $t$  wage schedule and a period  $t + 1$  price schedule at the very beginning of period  $t$ , before  $x_t$  is realized. This timing provides the maximal scope for insurance.

Firms must also coordinate output and employment levels across their commodity and labor agreements. We assume that firms meet all *ex post* demand given their pricing policy, so that demand determines employment. Hence fluctuations in demand generate movements in output and employment. The magnitude of this correlation depends upon the sensitivity of prices to past and present changes in the stock of money.

Consider a game between firms (both within and across generations) where the strategic variables are wage and price schedules. Workers and risk-averse consumers are nonstrategic players who exchange with a firm offering the most attractive wage and price schedules. The government and old age entrepreneurs purchase equal amounts from those firms that charge the lowest expected price. An equilibrium is then a sequence of wage and price functions  $(w_t, p_t)$ , dependent on events up to and including period  $t$ , which constitute a symmetric Nash equi-

librium in the game described above. Output in such an equilibrium is bounded by the productive capacity of the economy. Cooper (1984) discusses the detailed conditions for such equilibria in a closely related model.

Here we explore the possibility of a Nash equilibrium with wages and prices independent of policy shocks. Since realizations of  $x_t$  are public information,  $w_t$  will generally be indexed if  $p_{t+1}$  reflects  $x_t$ . Hence, we are searching for an equilibrium where both  $w_t$  and  $p_{t+1}$  are independent of  $x_t$ . Consider  $w_t = \alpha M_{t-1}$  and  $p_{t+1} = \beta M_{t-1}$ , for all  $t$ , as a candidate equilibrium. If  $\alpha/\beta = k$ , workers receive full insurance. The parameters of the contract can be set so that the expected consumption of a generation  $t$  entrepreneur satisfies  $E_{t-2} c_{t+1}^F = \bar{e}$  for all  $t$ . We expect this to emerge from the competition among firms.

A key issue here, as in the previous section, is the feasibility of such an equilibrium. Since quantities are demand determined, total employment (and hence output) in period  $t$  equals  $M_t/p_t$ . With  $p_t = \beta M_{t-1}$ , the maximal level of possible demand is  $(\bar{x})^2/\beta$ . Since the productive capacity of the economy is  $N\bar{n}$ , this predetermined wage/price equilibrium will exist iff  $N\bar{n} \geq (\bar{x})^2/\beta$ , where the parameter  $\beta$  turns out to be independent of  $\bar{x}$  and  $N\bar{n}$ . This condition reinforces our intuition that predetermined wage and price equilibria will not exist if policy is too "variable."

## REFERENCES

- Akerlof, George and Miyazaki, Hajime, "The Implicit Contract Theory of Unemployment Meets the Wage Bill Argument," *Review of Economic Studies*, January 1980, 47, 321-38.
- Azariadis, Costas, "Implicit Contracts and Underemployment Equilibria," *Journal of Political Economy*, December 1975, 83, 1183-202.
- Baily, Martin N., "Wages and Employment under Uncertain Demand," *Review of Economic Studies*, January 1974, 41, 37-50.
- Barro, Robert, "Long-Term Contracting, Sticky Prices and Monetary Policy," *Journal of Monetary Economics*, July 1977, 3, 305-16.
- Clower, Robert W., "The Keynesian Counter-Revolution: A Theoretical Appraisal," 1965; reprinted in D. Walker, ed., *Money and Markets*, Cambridge: Cambridge University Press, 1984.
- Cooper, Russell, "Expansionary Government Policy in an Economy with Commodity and Labor Contracts," Cowles Discussion Paper No. 727, October 1984.
- Farmer, Roger E. A., and Woodford, M., "Self-Fulfilling Prophecies and the Business Cycle," Working Paper, University of Pennsylvania, April, 1984.
- Fischer, Stanley, "Long-Term Contracts, Rational Expectations, and the Optimal Money Supply Rule," *Journal of Political Economy*, February 1977, 85, 191-205.
- Hart, Oliver, "Optimal Labour Contracts under Asymmetric Information: An Introduction," *Review of Economic Studies*, January 1983, 50, 3-36.
- Lucas, Robert E., "Expectations and The Neutrality of Money," *Journal of Economic Theory*, April 1972, 4, 103-24.
- Malinvaud, Edmond, *The Theory of Unemployment Reconsidered*, Oxford: Blackwell, 1977.
- Okun, Arthur, *Prices and Quantities*, Washington: The Brookings Institution, 1981.
- Sargent, Thomas J., and Wallace, Neil, "Rational Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule," *Journal of Political Economy*, April 1975, 83, 241-54.

# Wage Flexibility in the United States: Lessons from the Past

By DANIEL J. B. MITCHELL\*

Recently, various macroeconomic arguments for promoting wage flexibility and/or gain-sharing plans (such as profit sharing) have been developed. This paper explores why—despite such arguments—wage flexibility declined historically and profit-sharing-type arrangements have not covered most workers.

## I. Past Wage Flexibility

In another paper (forthcoming), I have contrasted wage setting in the 1920's with that of the post-World War II period. During the 1920's and early 1930's, the U.S. Bureau of Labor Statistics published an incomplete sample of reported wage-change decisions at the establishment level. Perhaps the best way to summarize the results is to direct attention to Table 1, which presents the distribution of manufacturing wage-change decisions during 1924 and 1925, years in which consumer price inflation was, respectively,  $-.2$  and  $+4.0$  percent on a December-to-December basis.

The table shows a wide array of wage-change decisions ranging from cuts of over 20 percent to increases of similar magnitude. This dispersion of decisions is remarkable by post-World War II standards. Moreover, the postwar evidence suggests that nominal wage cuts are a rarity, even in periods of low inflation. When they do occur, as in some recent union concessions, the cuts result from a painful negotiations process against a background of threatened or actual mass layoffs.

By the 1920's, many features of modern corporate enterprise were present. But unions were of little significance in most sectors, including manufacturing, the result of a sustained "open shop" campaign by employers

after World War I. There was little labor market intervention by government. Workers resented wage cuts—during periods of generalized wage cutting such reductions became important causes of strikes—but employers implemented them anyway. And when employers did not want to take the blame for wage cuts, they used "company unions" to "negotiate" reductions (Robert Dunn, 1927, pp. 21–23).

In short, in the absence of unions or other institutional constraints, the implicit contracts offered by employers in the 1920's provided substantially more wage flexibility than existed after World War II. The wage-setting mechanisms of the 1920's did not approach the flexibility of a classical auction market, a fact of some comfort to implicit-contract theorists. However, it is unclear that one needs to go much beyond simple explanations of how wage cuts (or even relative wage slippage) would lead to worker resentment and management caution.

## II. Why Did Wage Flexibility Diminish?

Implicit contract theories have emphasized worker risk aversion and/or turnover costs to explain wage unresponsiveness. Perhaps risk aversion and turnover costs increased after the 1920's, but the most significant changes that occurred involved unionization, new public policies, and changing social expectations.

In an important forthcoming book, Sanford Jacoby documents the development of internal personnel policy of American employers during the first half of this century. Jacoby points out that until World War I, employers were content to leave personnel policy to foremen. But during World War I, the budding field of personnel management—linked to other social welfare movements—took root. Employers created personnel departments to handle industrial relations on a centralized basis and to engage in "welfare work" among their employees. Their impetus

\*Graduate School of Management, University of California, Los Angeles, CA 90024.

TABLE 1—DISTRIBUTION OF MANUFACTURING WAGE-CHANGE DECISIONS

Percent Wage Change	1924	1925
-20 or less	2	1
-19.9 to -16	<sup>a</sup>	<sup>a</sup>
-15.9 to -12	4	2
-11.9 to -8	37	18
-7.9 to -4	8	6
-3.9 to -.1	4	3
.1 to 4	4	11
4.1 to 8	21	26
8.1 to 12	14	25
12.1 to 16	2	4
16.1 to 20	2	2
Over 20	1	2
All Decisions	100	100

Source: U.S. Bureau of Labor Statistics. See my paper (forthcoming) for detailed references.

Note: Shown in percent. Details need not sum to totals due to rounding. Unit of observation is the establishment.

<sup>a</sup>Less than 0.5 percent.

was a combination of war-related labor shortages and rapid growth in unionization which reflected various government policies aimed at achieving industrial peace and uninterrupted war production.

During the 1920's, in contrast, the labor shortage ended and government withdrew from the labor market. Unions were defeated as a threat. Firms downgraded their personnel departments, returned authority to foremen, and lost interest in workplace changes advocated by softminded reformers. When the union threat (supported by New Deal policies and later World War II controls) reappeared, the process reversed. Firms upgraded their personnel departments, downgraded foremen, and undertook various "progressive" personnel policies.

With the Great Depression came a purchasing power theory of wages that had a profound effect on public policy. Under this theory, depressions occurred if wages received too small a share of income. Low wages led to underconsumption. Maintaining wages, not cutting them, was seen as the appropriate policy for business downturns.

President Hoover urged firms not to cut wages at the onset of the depression (National Industrial Conference Board, 1932,

p. 3). And until 1931, his injunction seemed to have some effect. Pre-New Deal legislation, such as the Davis-Bacon Act (1931), reflected the wage-underconsumption idea. But the New Deal itself represented the high point of this theory, a theory supported by prominent economists of the time such as Paul Douglas (1935, pp. 212; 223-26), later a president of the American Economic Association.

The centerpiece of the early New Deal was the National Industrial Recovery Act (1933) that sought to boost business confidence by applying cartel-type codes on an industry basis. Code labor provisions set minimum wages, provided guides to wage differentials, and regulated other workplace conditions. Collective bargaining was also encouraged. At the bottom of a severe depression, a rash of wage *increases* suddenly developed as the result of these codes and wage cuts became the exception.

Although the NIRA was declared unconstitutional in 1935, elements were retained in such laws as the Wagner Act of 1935 (that contains the wage-underconsumption theory in its preamble) and the Fair Labor Standards Act (1938). Unemployment insurance (*UI*), established by the Social Security Act of 1935, is premised on layoffs rather than wage cuts as the response to falling demand. And since *UI* is not fully experience rated, it contains a net subsidy to layoff adjustments. Wage cuts can also have an adverse effect on future Social Security benefits received by workers. And they can adversely affect benefits to be paid under private (tax-code supported) pension plans.

More generally, the social ethos was changed by the depression. "Good" employers did not cut wages. To do so invited employee resentment and, possibly, unionization. Layoffs by seniority, tightly enforced in the union sector, less rigidly followed but still influential in the nonunion sector, became the basic mode of adjustment. And lest employers slip, they now have the Age Discrimination in Employment Act (1967, 1978)—age and tenure are highly correlated—to remind them to respect seniority.

After the 1940's, government programs created a "third wave" of pressures strengthening centralized employer personnel poli-

cies. The tax code was amended to favor a variety of employer-provided fringe benefits requiring expertise in administration. In the 1960's various federal and state "equal employment opportunity" programs further elevated the control of personnel departments. And the 1970's saw new legislation in occupational safety and health, pension plan regulation, and other areas which continued the trend. Centralized personnel bureaucracies are not conducive to flexible wage decisions.

In short, there is considerable evidence that what changed after the 1920's was first the union threat, and second, a variety of public policies and expectations. Modern unionized firms exhibit wage unresponsiveness because of strategic factors (for example, strike costs) involved in negotiations—that lead to averaging out the business cycle under long-term contracts—and because "median voter" tastes and seniority systems lead to a preference for layoffs rather than wage cuts. Nonunion firms have historically been concerned about the union threat and have evolved strong personnel bureaucracies to administer "unionese" employee-welfare programs.

### III. Why Don't We Have More Profit Sharing?

The tendency toward unresponsive wage rates could be overcome by gain-sharing arrangements such as profit sharing. But if profit sharing has macroeconomic benefits, it will be used less than is socially optimal since employers (and unions) will not take account of the externalities. Certainly, current wage practices are far removed from a gain-sharing economy.

Although externalities will be ignored, employers (and unions) might establish profit sharing based on private incentives. Why haven't these incentives inspired more widespread use of profit sharing? The answer cannot be that the technology is new. Profit sharing reportedly began in France in the 1840's; then spread to other countries (Lyle Cooper, 1934). The first profit-sharing plan in the United States appeared in 1867; by 1916, the Bureau of Labor Statistics undertook a study of profit-sharing plans and

found a total of sixty in operation covering about 30,000 workers (Boris Emmet, 1917). Profit sharing provoked academic interest as early as 1887, when a debate on the issue appeared in the *Quarterly Journal of Economics* (Richard Aldrich; Franklin Giddings), and continued into the 1920's (F. E. Wolfe).

There were three initial motivations for installing profit sharing. Some employers saw such plans in moral or religious terms; one characterized his plan as an application of "divine law" which "ushered in love, contentment, cooperation, and happiness (and) cast out hell..." (1920, U.S. Bureau of Labor Statistics). However, such employers were rare. More commonly, employers saw profit sharing as a way of building employee loyalty, thus avoiding industrial unrest and unions. Finally, profit sharing was advocated as a way of putting the employee on the side of management, thereby boosting production and efficiency (Daniel Bloomfield, 1923, p. 59).

For employers interested in building cooperative, loyal, nonunion workforces, various alternatives to profit sharing were available. Company unions and employee representation plans could be created. During World War I, the federal government fostered such "works councils," some of which persisted into the 1920's and beyond. These arrangements sometimes had "collective economy dividend" plans attached as a form of gain sharing (National Industrial Conference Board, 1922, p. 2). But employers often found that profit sharing was not sufficiently appreciated or understood by workers.

It is not surprising that profit sharing came to be unpopular with unions (1916, American Federation of Labor). Where it was used, it was often billed as an anti-union tool. Terminology was loose; employers often referred to paternalistic bonus arrangements as "profit sharing" even if they bore no relation to profits. At Ford, for example, a bonus unrelated to profits—but termed profit sharing—was distributed to workers who met company standards of morality as determined through home visits by company investigators. Union antipathy toward profit sharing was well ensconced by the late 1940's (Kenneth Thompson, 1949, pp. 43–55) and



did not appreciably shift until the recent concession bargains.

As an incentive system, profit sharing competed against a plethora of alternative arrangements. The 1920's was a period of widespread use of piece rates and other incentive plans, reflecting the earlier development of "scientific management." Roughly one-half of production workers were covered by incentive rates and bonuses. In contrast, a survey in the early 1960's found the fraction at about one-fourth (see my forthcoming article). These incentive programs were closely related to individual or group effort. But profit sharing was remote from the effort of the individual worker. Moreover, profits were affected by influences other than worker effort, so that bonuses appeared unrelated to employee achievement (Gordon Watkins, 1922, p. 517).

Thus, profit sharing had difficulty competing, either as a loyalty generator or an incentive system. It initially received unfavorable tax treatment; profit-sharing bonuses were considered non-expensable "gratuities" (Emmet, p. 6). Later, the variable compensation element of profit sharing became the antithesis of the New Deal's emphasis on income security and stability.

After World War II, unions pushed for more income stability through a "guaranteed annual wage." One union did suggest that a profit-sharing fund could be used to finance such a program (Arnold Frutkin and Donald Farwell, 1955, p. 133). But the eventual outcome was a fixed employer contribution to a supplemental unemployment benefit fund rather than a profit-related payment.

#### IV. Current Opportunities

If profit-sharing plans, or other forms of gain sharing, have desirable, external macroeconomic effects, a case can be made for subsidizing their use. At present, the tax code offers favorable treatment only to profit-sharing plans which place their bonuses into deferred (i.e., retirement) funds. Cash distributions receive no advantageous treatment. Moreover, qualified plans need not specify a fixed formula by which the bonus is determined. Thus, many profit-sharing plans

are ersatz pensions that escape the more rigorous government regulation applied to formalized pension programs.

Historically, employers who were persuaded that giving their employees a "stake" in the firm's welfare would contribute to worker loyalty, often chose various forms of employee stock ownership plans (*ESOP*) instead of profit sharing. Except in the extreme cases in which the workers own all the shares, such arrangements will not have the desirable macroeconomic effects of gain sharing. Congress, however, has been persuaded to bestow favorable tax treatment on *ESOPs*, in response to "every-worker-a-capitalist" arguments.

It is not just incentives (such as risk aversion and turnover costs) that determine firm wage policies; changes in institutions and social expectations also play a role. The wave of union wage concessions since 1979 has created an historical opportunity to promote gain sharing. Although most concessions have not involved installation of profit or other forms of gain sharing, some prominent bargains in autos, airlines, and other industries have moved in this direction. Tax incentives tailored to those forms of gain sharing which provide macroeconomic benefits—and other forms of promotion of gain sharing—could take advantage of the change in climate and promote a more stable economy.

#### REFERENCES

- Aldrich, Richard, "Some Objections to Profit Sharing," *Quarterly Journal of Economics*, January 1887, 1, 232-42.
- Bloomfield, Daniel, *Financial Incentives for Employees and Executives*, Vol. 11, New York: H. W. Wilson, 1923, 58-127.
- Cooper, Lyle W., "Profit Sharing," in R. A. Seligman and Alvin Johnson, eds., *Encyclopedia of the Social Sciences*, New York: MacMillan, 1934, 487-92.
- Douglas, Paul H., *Controlling Depressions*, New York: W. W. Norton, 1935.
- Dunn, Robert W., *Company Unions*, New York: Vanguard Press, 1927.
- Emmet, Boris, *Profit Sharing in the United States*, BLS Bulletin 208, Washington:

- USGPO, 1917.
- Frutkin, Arnold W. and Farwell, Donald F., *The Guaranteed Annual Wage*, Washington: Bureau of National Affairs, Inc., 1955.
- Giddings, Franklin H., "The Theory of Profit Sharing," *Quarterly Journal of Economics*, January 1887, 1, 367-76.
- Jacoby, Sanford M., *Employing Bureaucracy: Managers, Unions, and the Transformation of Work in American Industry*, New York: Columbia University Press, forthcoming 1985.
- Mitchell, Daniel J. B., "Wage Flexibility; Then and Now," *Industrial Relations*, forthcoming.
- Thompson, Kenneth M., *Profit Sharing: Democratic Capitalism in American Industry*, New York: Harper & Brothers, 1949.
- Watkins, Gordon S., *An Introduction to the Study of Labor Problems*, New York: Thomas Y. Crowell, 1922.
- Wolfe, F. E., "A Survey of Profit-Sharing and Bonuses in Chicago Printing-Plants," *Journal of Political Economy*, July 1921, 29, 521-42.
- American Federation of Labor, "The Stetson Strike and Profit-Sharing," *American Federationist*, May 1916, 23, 383-85.
- National Industrial Conference Board, *Experience with Works Councils in the United States*, New York: Century, 1922.
- \_\_\_\_\_, *Salary and Wage Policy in the Depression*, New York: National Industrial Conference Board, 1932.
- U.S. Bureau of Labor Statistics, "Application of the Golden Rule in Business," *Monthly Labor Review*, December 1920, 11, 1222-23.

# Profit Sharing as Macroeconomic Policy

By MARTIN L. WEITZMAN\*

When Keynes came to sum up the central message of the *General Theory* for the economics profession, in a remarkable but by now long-forgotten *QJE* article of 1937, he began with a "general, philosophical disquisition on the behavior of mankind"—under uncertainty. Here as elsewhere, Keynes made it abundantly clear that he shared Frank Knight's distinction. "Uncertainty" did not mean "risk"—that which is, at least in principle, reducible to well-defined actuarial probabilities. By uncertainty Keynes intended, I believe, to convey the idea of "ignorance"—that which is essentially due to insufficient or precarious knowledge of the mechanism by which the future is generated out of the past.

The Keynesian scenario looks out over an economic world that is rife with uncertainty. In that world, expectations play an important dual role as both a manifestation of uncertainty and a cause of it. Such expectations are arbitrary to some degree because they can be based on almost anything, including self-fulfilling expectations of the behavior and expectations of others. And, as Keynes pointed out, "being based on so flimsy a foundation," these expectations of expectations are "subject to sudden and violent changes."

It follows that while there may ultimately be some long-run forces drawing it toward full employment, capitalism may also have some deep-seated tendencies toward short-run instability. Unadulterated *laissez-faire* is likely to be out of equilibrium much of the time, and even when it is in equilibrium there is no guarantee of being in a "good" equilibrium. Whether in a state of "bad" equilibrium or merely in disequilibrium, such coordination failures generate undesirable macroeconomic consequences like unem-

ployment which can cause very significant welfare losses. By the ultimate logic of this Keynesian worldview, then, the stage is set for some form of government intervention to recoordinate the economy into a better configuration. Any such government policy will inevitably introduce some microeconomic distortions, but as an empirical matter such losses tend to be small, relative to the enormous welfare gains from having an economy operate at its full-employment level.

Such general considerations do not indicate the best *form* of government intervention to stabilize the macroeconomy. Indeed, we do not currently have a general, realistic framework within which a meta-issue like that might be properly addressed. Nevertheless it is possible, I believe, to give some common sense criteria for desirable forms of government intervention. It is my contention that economists have not been sufficiently imaginative in devising operational mechanisms or systems possessing advantageous macroeconomic properties. The usual fiscal and monetary policies are, to my mind, sledgehammer-like tactics for controlling unemployment and inflation. They do the job, but clumsily, by brute force—and they can leave a big mess afterwards. I think it is possible to find subtler alternatives that operate more cleanly and with a softer touch by taking a page from the book of Adam Smith.

A good mechanism for fighting unemployment and inflation should have several noteworthy characteristics. It should be decentralized, based on the natural microeconomic incentives of a market-like environment. It should work more or less automatically, keeping to a minimum the need for using discretionary government policy. And, in a highly uncertain world, it should be robust in retaining its desirable macroeconomic characteristics over a wide range of possible situations or circumstances—including some that are currently unforeseen.

I want to argue that a superior form of government policy for combating unemploy-

\*Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139.

ment and inflation is to encourage, through exhortation and special tax privileges, the widespread use of profit sharing. A profit-sharing system has the potential to automatically counteract contractionary or inflationary shocks—while maintaining the advantages of decentralized decision making. And these desirable properties are robustly preserved throughout a variety of economic environments. At the very least, widespread profit sharing can be a valuable adjunct to traditional monetary and fiscal policies.

I believe we should seriously consider some new ideas about basic reform of the economic mechanism because our old ways of doing things are no longer adequate. The premier economic malady of our time is stagflation. Despite some abatement of its virulence in the immediate present, we still seem to be unable to reconcile, over a reasonably sustained period, high employment with low inflation. Even when economic conditions are on the upswing, significant pockets of unemployed workers remain throughout the Western capitalist countries. Right now, for example, we are afraid to aggressively push unemployment down to more humane levels for fear of re-igniting inflation. The policy-induced recession remains our only reliable method for lowering inflation rates. It is difficult to imagine a more costly, inefficient, or unjust waste of economic resources and human potential. Profit sharing represents a way of building into the system the kind of natural resistance to unemployment and inflation that could really disarm stagflation at its source.

Our macroeconomic problems trace back, ultimately, to the wage system of paying labor. We try to award every employed worker a predetermined piece of the income pie before it is out of the oven, before the size of the pie is even known. Our "social contract" promises workers a fixed wage independent of the health of their company, while the company chooses the employment level. That stabilizes the money income of whomever is hired, but only at the considerable cost of loading unemployment on low-seniority workers and inflation on everybody—a socially inferior risk-sharing arrangement that both diminishes and makes more

variable the real income of the working class as a whole.

Why does a profit-sharing system possess superior macroeconomic properties that help to automatically stabilize output at the full-employment level and make it easier to deal with inflation? There is not sufficient space in this condensed paper to give a detailed answer, so the true seeker must be prepared to fight through the longer and more technical pieces listed as references (1983, 1984). But a shorter heuristic story, a kind of summary, can be briefly told here.

Consider a typical monopolistically competitive firm in a partial equilibrium setting. Suppose the wage is treated as a quasi-fixed parameter in the short run. If the firm can hire as much labor as it wants, it will employ workers to the point where the marginal revenue product of labor equals the wage rate. This is familiar enough. Consider, though, what happens with a profit-sharing contract that names a base wage and a certain fraction of profits per worker to be paid to each worker. Suppose these two pay parameters are treated as quasi fixed in the short run. A little reflection reveals that if the profit-sharing firm can hire as much labor as it wants, it will employ workers to the point where the marginal revenue product of labor equals the base wage, independent of the value of the profit-sharing parameter. (Note, though, that what the worker is actually paid depends very much on the value of the profit-sharing coefficient.) When a standard *IS-LM*-type macro model is constructed around such a model of the firm, the following isomorphism emerges. A profit-sharing macroeconomy will find itself with the same output, employment, and price level as the corresponding wage economy whose wage is set at the profit-sharing economy's base-wage level. In other words, the aggregate macroeconomic characteristics of a profit-sharing economy, excepting the distribution of income, are determined (on the cost side) by its base wage alone. The profit-sharing parameter does not influence output, employment, or prices, although it does influence the distribution of income. If the employed workers can be persuaded to take more of their income in the form of profit shares and less in

the form of base wages, that can result in a Pareto improvement—with increased aggregate output and employment, lower prices, and higher real pay.

When identical-twin wage and profit-sharing economies are placed in the same stationary environment, with competitive labor markets, both economies will gravitate toward the same long-run full-employment equilibrium. But, then perform the following thought experiment. In the typical style of disequilibrium analysis, disturb each economy and observe the short-run reaction when pay parameters are quasi fixed but everything else is allowed to vary. The profit-sharing economy will remain at full employment after a disturbance, while a contractionary shock will cause a wage economy to disemploy labor. It should not be hard to imagine why such characteristics make a profit-sharing system more resistant to stagflation.

This same point can be made yet another way. Consider the standard textbook *IS-LM*-type model. Aggregate demand is determined, via the appropriate multipliers, as a function of autonomous spending injections and real money balances. The price level is determined as a degree-of-monopoly-power markup over wages. Wages are treated as exogenously fixed in the short run. Given the standard *IS-LM*-type specification, the model grinds out (as a parametric function of the wage level) output, employment, and the price level. It is clear what happens within such a model if there is a *ceteris paribus* money-wage cut. Output and employment are higher, while prices are lower. Yet this is exactly what occurs when an economy shifts toward profit sharing. The base wage determines the fundamental macroeconomic characteristics of the system—when there is an increase in profit shares at the expense of base wages, macroeconomic performance improves without loss of real labor income.

I am aware that such short-run, fixed-pay-parameter, disequilibrium models will be unsatisfying to the economic theory purist who will want a full-blown account of why one payment mechanism rather than another has been selected by society in the first place, and who will not rest content without under-

standing on a more fundamental level why pay parameters should be sticky in the short run. Such concerns have a legitimate place. But I do not think they should be taken to such an extreme that we are inhibited from examining what would happen in disequilibrium under alternative payment systems before first having firmly in hand a general, all-encompassing theory of economic systems and disequilibrium-like behavior.

What about the possible objections to profit sharing? Several are frequently voiced. I believe the objections can be successfully rebutted, but here I deal with only one, and that skimpily. The objection to profit sharing one hears most often from economists is that, compared with a wage system, it represents a socially inefficient method of risk sharing. (Isn't it obvious that under a wage system, the firm bears the risk, while under a profit-sharing system, the worker bears the risk?) In my opinion the reasoning traditionally put forward to support this "insurance" argument is fallacious, being based on a partial equilibrium view that does not take into account the radically different macroeconomic consequences of the two systems for overall employment and aggregate output. The fixed wage does not stabilize labor income. What is true for the individual tenured worker is not true for labor as a whole. When a more complete analysis is performed, which considers the situation not as seen by a tenured, high-seniority worker who already has job security, but by a neutral observer with a reasonably specified social welfare function defined over the entire population, it becomes clear that the welfare advantages of a profit-sharing system (that delivers permanent full employment) are enormously greater than a wage system (that permits unemployment). The basic reason is not difficult to understand. A wage system allows huge first-order Okun-gap losses of output and welfare to open up when a significant slice of the national income pie evaporates. A profit-sharing system stabilizes aggregate output at the full-employment level, creating the biggest possible national income pie, while permitting only small second-order Harberger-triangle losses to arise because some crumbs have been randomly

redistributed from a worker in one firm to a worker in another. Here is a friendly challenge to would-be critics. I challenge anyone to cook up an empirical real world scenario, with reasonable numbers and specifications, where a profit-sharing system does not deliver significantly greater social welfare than a wage system.

The superior profit-sharing variant of capitalism is practiced, to some extent, in the immensely successful economies of Japan, Korea, and Taiwan. While these countries are not identical clones, their economies do share certain important characteristics. In each case workers receive a significant fraction of their pay in the form of a bonus. The bonuses are large, averaging over good years and bad about 25 percent of a worker's total pay in Japan, and about 15 percent in Korea and Taiwan. The degree to which the bonus is actually determined as a function of current profits per worker varies from firm to firm, and depends upon the country. (For example, in some Japanese companies the bonus is almost a disguised wage, but this is not true for most Japanese companies, and it appears to be hardly true for any Korean companies.) Bonuses, like dividends, respond to corporate earnings, but with a complicated lag structure not easy to quantify by any rigidly prescribed rule. Overall, there is very little question that profit sharing is a significant feature of the industrial landscapes of these "Japanese-style" economies.

While it is difficult to quantify the exact magnitude of its contribution out of a host of reinforcing tendencies, the bonus system is almost surely one major reason (although, most likely, far from the only reason) for the outstanding economic performances of Japan, Korea, and Taiwan. Their flexible payment system helps these economies to ride out the business cycle with relatively high, stable levels of employment and output. Their governments enjoy greater leeway for fighting inflation without causing unemployment. The variability of real pay per member of the potential labor force has actually been reduced. Over time, a more equitable distribution of income has emerged than is found in other capitalist countries.

I believe that we in the West, instead of giving lessons as we are accustomed to doing, now must be prepared to take a lesson from the East. We should consciously tilt our economies toward this superior variant of capitalism. We ought to adopt a new social contract that promises our working people full employment without inflation but asks, in return, that workers receive a significant fraction of their pay in the form of a profit-sharing bonus.

But, the typical economist will ask, why if a profit-sharing system represents a far better way of operating a market economy than a wage system don't we see more examples of share economies? After all, even in Japan, Korea, and Taiwan only modest (although significant) steps have been taken in this direction. The rest of the advanced capitalist countries are predominantly wage economies. Why, if profit sharing is so beneficial, does not self-interest automatically lead firms and workers in this direction?

The answer involves an externality or market failure of enormous magnitude. In choosing a particular contract form, the firm and its workers only calculate the effects on themselves. They take no account whatsoever of the possible effects on the rest of the economy. When a firm and its workers select a labor contract with a strong profit-sharing component, they are contributing to an atmosphere of full employment and brisk aggregate demand without inflation because the firm is then more willing to hire new "outsider" workers and to expand output by riding down its demand curve, lowering its price. But these macroeconomic advantages to the outsiders do not properly accrue to those insiders who make the decision. Like clean air, the benefits are spread throughout the community. The wage firm and its workers do not have the proper incentives to cease polluting the macroeconomic environment by converting to a share contract. The essence of the public good aspect of the problem is that, in choosing between contract forms, the firm and its workers do not take into account the employment effects on the labor market as a whole and the consequent spending implications for aggregate

demand. The macroeconomic externality of a tight labor market is helped by a share contract and hurt by a wage contract, but the difference is uncompensated. In such situations there can be no presumption that the economy is optimally organized and society-wide reform may be needed to nudge firms and workers towards increased profit sharing.

This much-needed reform will not come about easily. Persuading workers and companies to change fundamentally the way labor is paid, in the name of the public interest, will demand political leadership of a very high order. Material incentives will probably be required, such as favorable tax treatment of the profit-sharing component of a worker's pay. Yet the benefits of full employment

without inflation are so enormous and the increased income is so great, that we cannot afford not to move in this direction.

#### REFERENCES

- Keynes, John Maynard, "The General Theory of Employment," *Quarterly Journal of Economics*, February 1937, 51, 209-23.
- Weitzman, Martin L., "Some Macroeconomic Implications of Alternative Compensation Systems," *Economic Journal*, December 1983, 93, 763-83.
- , *The Share Economy*, Cambridge: Harvard University Press, 1984.
- , "The Simple Macroeconomics of Profit Sharing," unpublished working paper, June 1984.

# THE CONSEQUENCES OF DEREGULATION IN THE TRANSPORTATION AND TELECOMMUNICATIONS SECTOR†

## The Regulatory Transition

By JOHN R. MEYER AND WILLIAM B. TYE\*

A number of regulated industries, particularly in transportation and communications, have recently undertaken the transition from a regime of rigid price and entry controls to that of a more competitive market structure. While the different industries have experienced somewhat different fates, the responses to this transition do have certain common underlying characteristics.

To start, demands for some form of temporary or continuing regulation during the transition to deregulation can be explained almost entirely as a response to the strength of the entry threat relative to the magnitude of sunk costs incurred by the affected parties in the previous regulatory regime. Where the obstacles to entry are low, the incumbent firms and labor ordinarily seek "protective conditions" during the transition to permit them to recover some or all of their sunk costs. When the obstacles to entry are high, customers are likely to make similar demands for protective conditions designed to do the same, particularly when the customers' own sunk costs severely restrict their competitive options after deregulation.

Pleas for protective conditions during the transition are widely regarded as introducing market imperfections that should be resisted in the name of regulatory reform. This view, however, naively equates the market results during the transition (when choices are constrained by the presence of sunk costs) to the results that would prevail in a long-run equilibrium where deregulated prices and quanti-

ties are established in the absence of most (or any) sunk costs. The regulatory problem during the transition is to define a set of residual (hopefully self-terminating) economic constraints that will satisfy the equity and other considerations created by the short- to medium-term continuation of some sunk costs without creating insurmountable obstacles to approaching an efficient competitive outcome in the long run.

Any transition mechanism must thus come to grips with the essence of the transition problem from a political as well as an economic perspective: who is to bear the consequences of the "overhang" of sunk costs. Note that we are not making a generalized plea for the compensation of losers from deregulation, especially for windfall gains conferred by the regulatory process itself (see Kenneth Gordon, 1981). Rather, the transition problem is defined here to be a limited period in which participants in the regulatory game are permitted to amortize financial commitments made under the prior set of rules while other participants are constrained in their ability to exploit the presence of those sunk costs during the transition. Misunderstanding or failing to recognize this transition problem can pose substantial dangers: specifically, premature application of economic concepts that, while arguably valid in some future regime in which all sunk costs are amortized, decidedly do not account for the effect of these sunk costs on the marketplace in the short run. Misunderstandings of the transition problem may also encourage false conclusions about the eventual results of deregulation, that is, the long-run competitive equilibrium and industry structure that will emerge. As a consequence, policy recommendations designed to address the problems of the transition may

†*Discussants:* Robert Willig, Princeton University; Thomas Moore, Hoover Institution.

\*Harvard University, Cambridge, MA 02138, and Putnam, Hayes & Bartlett, Inc., 124 Mt. Auburn Street, Cambridge, MA 02138, respectively.



inadvertently frustrate the ultimate goal of deregulation.

The transition problem arising from sunk costs is exacerbated by the fact that deregulation also undermines the prior "regulatory equilibrium." That equilibrium, even though it might bear little resemblance to a true market equilibrium, usually did manage to apportion the total costs of the regulated enterprise in a fashion that more or less kept the factors of production committed. Typically, regulators instituted what amounted to informal taxation schemes. Under these schemes, different consumers, sometimes aided by government subsidies, paid the needed revenues but individual charges often bore little resemblance to the individual costs of the services performed, so that "contributions" could vary widely from one activity to another. Finding a "revenue-adequate" substitute for the old regulatory excise tax scheme during the transition to deregulation can be a difficult task, giving rise to numerous tradeoffs and conflicts, especially if combined with efforts to protect certain so-called captive or otherwise highly dependent classes of customers.

All these problems of defining revenue adequacy and achieving it without unleashing too many political charges of unfairness are perhaps best illustrated by the railroad industry, now into its fifth year of regulatory reform under the *Staggers Rail Act*. According to the *Staggers Act*, implementation of national transportation policy should start with the presumption that 1) an effectively competitive sphere and a (largely transitory) "market-dominant" sphere coexist in the rail industry; 2) regulation over rate reasonableness is to be maintained only in markets where there is an absence of effective competition; and 3) competition should be nurtured and encouraged in order to maximize the size of the competitive sphere and thus minimize the need for residual economic regulation in the market-dominant sphere. Thus, in the *Staggers Act*, Congress established a so-called reasonable "base rate" (which could be escalated for inflation and revenue-adequacy needs) for captive shippers where the Interstate Commerce Commission (ICC) found the presence of market dominance. All

other existing rates (competitive shippers) were essentially deregulated insofar as rate reasonableness was concerned. In essence, Congress was responding to shippers who feared that they would be charged "unreasonable rates" during the regulatory transition because they had sunk substantial costs which made them captive to a single carrier, for example, a coal-fired electric generating plant with a remaining lifetime extending over several decades consuming coal purchased under long-term contract from a specific mine served by only one railroad and with no viable barge or other transport alternatives.

After the passage of the *Staggers Act*, numerous economists (William Baumol, Robert Willig, and Stephen Goldfeld, 1981; and Mark Levin and Bruce Stram, 1981) endorsed Ramsey Pricing<sup>1</sup> as the economically rational standard for setting rail rates in the captive sector. Many shippers, though, wondered how a rate design predicated on assigning the highest rates to captive shippers (indeed, whatever it took for the railroad to become revenue-adequate, subject only to the shippers' unaided search for alternatives) qualified as what Congress had in mind as a transition device to protect captive shippers under the *Staggers Act*. Apparently in response to this criticism, the ICC moved to constrain the results of Ramsey Pricing with a so-called "surrogate for competition," the Stand-Alone Cost (SAC) test (ICC, 1983) which "...in brief, requires that the consumer of a service be charged a price no higher than that at which it could be offered by a *specialized* competitive supplier" (Baumol and Willig, 1983, p. 7; emphasis added).

Critics of Ramsey Pricing and the SAC test have argued that these concepts are not

<sup>1</sup> Ramsey Pricing, a variant of value-of-service pricing or the Inverse Elasticity Rule, sets the highest rates for the most demand-inelastic traffic (E. P. A. Ramsey, 1927; Harold Hotelling, 1938; Baumol et al., 1962; and Baumol-David Bradford, 1970). In the context of railroad ratemaking, Ramsey Pricing (and its economic rationale) was largely anticipated in the concept of value of service ratemaking and the writings of D. Phillip Locklin (1933).

responsive to Congressional intentions to protect captive shippers (see Tye, 1983a,b) nor are they practical to implement (Tye and Herman Leonard, 1983; and Tye, 1985). In particular, they object that the SAC test is a very poor surrogate for real competition. A true "surrogate," they argue, must look to *all* the constraints on market power in a competitive market—intramodal, intermodal, product, and geographic. The *specialized* SAC test is a poor surrogate for the richness of this competitive environment because it only looks at a very narrow kind of potential competition, that of a specialized entrant who by definition could never flourish in the purported environment, that of strong economies of scope. Such economies arise if the total costs to a carrier of supplying multiple services simultaneously is less than the sum of the costs of supplying each of them separately. (See Willig, 1983; and John Panzar-Willig, 1977.) Joint costs are one origin of economies of scope (see Panzar-Willig, 1981). These critics therefore view the SAC test as a necessary but hardly sufficient surrogate for effective competition and not the only reliable measure of a maximum reasonable rate during the transition.

Proponents of Ramsey Pricing and the SAC test never explicitly define their vision of the equilibrium industry structure after the transition, but it can be inferred from their assumptions (implicit and explicit) about cost structure and their policy prescriptions for the transition. Their vision seems to be a rail industry with allegedly marked economies of scope and scale where average prices must be substantially above long-run marginal cost in order to achieve revenue adequacy (Tye, 1984). With such a vision of the deregulated equilibrium, Ramsey Pricing (reined in only by the SAC test) during the transition could well become mainly a redistributive device, transferring much of the burden of unwinding unneeded investments in rail capacity onto certain captive shippers, as contrasted, say, with placing that burden strictly on rail labor and capital. This obviously raises equity issues that seem better suited to political than economic adjudication.

Furthermore, discriminatory prices that cannot be sustained in the long run because

of broader (nonintramodal) competition may be self-defeating for the railroads (and therefore presumably avoided by management with perfect foresight and complete control). Specifically, if a railroad is economically viable in the long run, it is not in the railroad's self-interest to undermine or relocate a shipper—especially a captive shipper—as long as that shipper pays rates that make a positive contribution to covering nonallocable costs. Even if the railroad contemplates going out of business (for example, abandoning service to a particular shipper), an optimal rate policy would take into account possible tradeoffs between maximizing current contribution and the rate at which the shipper went out of business; that is, a lower rate today might yield a larger discounted net value of contributions. Another, somewhat opposite danger for carriers practicing price discrimination is that they may further sink costs during the regulatory transition (based on the mistaken belief that highly discriminatory pricing structures are a permanent feature of a competitive market), only to discover that their captive shippers have slipped from their grasp, once again leaving them revenue-inadequate.<sup>2</sup>

Lower rates in markets temporarily glutted by overcapacity may also not be an unmitigated benefit to the shipper or consumer, especially if the transient character of the low rates is not understood. Again, if the low rates are mistakenly expected to persist, "downstream" investments or locations may be incorrectly chosen by shippers. Indeed, this is essentially the "mistake" made by many of today's captive shippers in that they incorrectly thought that regulation would perpetuate the historical rate patterns. Of course, the converse could also occur: excep-

<sup>2</sup>Proposals to pursue more short-run or discriminatory pricing schemes in the rail industry have been strongly endorsed by practitioners of the theory of "contestable markets," with its strong emphasis on the role of sellers' sunk costs. This endorsement is paradoxical, given that Ramsey Pricing implies overlooking buyers' sunk costs for pricing policy during the regulatory transition. Such a system of essentially unfettered price discrimination can be enforced only with the constraints on buyers' choices that such sunk costs temporarily provide.

tionally high rates assessed on today's captive shipper might deter investments or reinvestments that might be mutually beneficial to both the investor and the carrier in the longer run. In short, substantial departures from prices based on long-run cost considerations, even though sometimes bestowing short-run benefits, have a good deal of potential for achieving misallocation of resources and other economic mischief in the long run. This is especially likely to be true in an economy characterized by rapid technological and demographic changes that are constantly opening up new economic possibilities and combinations.<sup>3</sup>

Problems of matching financial commitments and risks, or of becoming captive to a single vendor, are, of course, not unique to the rail industry. They often arise when a long-lived investment is to be used by multiple tenants. Once committed to such a project, joint costs are substantially sunk and the tenants have closed off other options. For that reason, there is no true market solution for allocating the joint costs among captive users once they are sunk. However, a market for long-term contracts could have accomplished these objectives in the absence of regulation. Given the risks of committing to large amounts of fixed capacity for long periods of time (for example, in bulk shipping, power plants, commercial office space, mining operations, oil pipelines, air terminals, etc.), investors usually size their installations to match long-term contractual commitments made by buyers, though sometimes reserving some extra capacity for the "spot" market or growth in demand (Meyer, 1983). Such contracts shift some of the risk to buyers who are presumably more knowledgeable about their future demands than sellers, and insulate sellers from the rate wars which inevitably break out on the spot markets when excess capacity persists over sufficiently long periods. Equally as important, such contracts permit customers to make decisions that leave them "captive" to suppliers; that is, they are then free to make

decisions to sink additional costs that substantially reduce their access to competing vendors.

Had the rail industry developed without regulatory oversight, such a "contracting equilibrium" might already be in place. Unfortunately, recreating that equilibrium retroactively is not easily done—for otherwise it would seem to be a good approximation to the *ex post* "surrogate for competition" that Congress and the regulatory commissions seek. Among other problems is how to treat *ex post* "excess capacity"; in concept, both shippers and carriers agree that shippers should pay only the cost of "needed" capacity, but that concept immediately invokes debate over how much capacity would be provided in a "freely contracted" competitive equilibrium. In general, much of the extant capacity in the industry might never have been constructed (or maintained) without prior contractual commitments from some shippers to pay substantially more than they can be assessed under an *ex post* deregulated scenario. A conflict once again arises between the *Staggers Act* objective of revenue adequacy, which would suggest making the captive shipper bear these sunk costs in the short run, and the *Staggers Act* guarantees to captive shippers, which suggests that the carrier bear them if competition precludes recovery from the competitive sector.

In a "contractual approach" to the regulatory transition the existence of barriers to entry and substantial sunk costs would also mean that prices would never be set by "hit-and-run" entry, as in more competitive or contestable markets. Above all, the contractual approach directs attention to the long-run choices made when buyers have few or no sunk costs, including all the competitive choices available in the long run—intermodal, intramodal, product, and geographic. Price discrimination arising from buyers' and sellers' sunk costs should not exist in such a "contractual equilibrium."

The resulting rate ceiling for captive shippers during the transition could, of course, be substantially less permissive than that offered by the SAC test. Such a policy, though, should have a better chance of maintaining the consensus for deregulation during the transition and would also represent a

<sup>3</sup>The rail industry has not been noted for its ability to foresee long-term trends and to change its pricing strategy in response (see Meyer et al., 1959).

clear signal to the carriers that the transition to a deregulated environment will be swift and that they will not be afforded the luxury of a slow adjustment at the expense of the captive shipper. In general, any deregulated industry should not be encouraged to use residual regulation during the transition process to seek additional ways to preclude efficient choices by shippers, choices they might accept freely under contract. Such a policy will only encourage those customers to seek out alternatives that do not use these recently deregulated services while perhaps tempting carriers and regulators to seek new ways of restricting competition in an effort to extend the transition period.

In short, two distinctly different regulatory transitions can be identified which raise different equity and welfare problems and require somewhat different policy responses. One situation is characterized by relatively easy entry and few commitments that tie particular consumers to particular vendors. The other is characterized by limited entry possibilities and some consumers who have sunk costs or made commitments that do tie them for at least some time to particular vendors. Of course, in the real world the two situations will rarely be encountered in pure forms but rather in mixes, so that mixed-policy strategies may also be advisable.

In the first situation, that of relatively free entry, economists' notions of workable competition or contestable markets, however labelled, will seem largely applicable and little or no residual regulation will be needed to protect consumer interests. Indeed, complaints about deregulation in such circumstances are more likely to originate with the factors of production—labor and capital—for whom deregulation can often mean erosion of a previously privileged position (i.e., capture of rents created by the regulatory process itself). But from the standpoint of consumer (or shipper) interests alone, a prompt, almost immediate, transition would seem advisable (Alfred Kahn, 1979).

By contrast, in the second situation the transition process can involve some aggrieved consumers who perceive themselves as unfairly victimized by the transition. The difficulty is created by an overhang of sunk

costs committed under the prior rules of regulation. Eventually, these problems will be resolved by the means usually employed in a market economy when long-term commitments must be made by different participants in an enterprise—the establishment of mutually satisfactory contractual agreements between the participants. Until then, however, public policy is likely to be confronted with some difficult decisions of equity, in which the different participants can make plausible claims about the justice or injustice of different policies. Designing an acceptable regulatory transition in these circumstances will almost certainly be difficult and time consuming.

Are there any economic guidelines for steering public policy through these transition difficulties and compromises? Above all, appropriate public policy during the transition should avoid trying to outguess the market by imposing regulatory actions designed to achieve preconceived visions of the long-run industry equilibrium. (In a world of rapid technological, demographic, and other changes, forecasting the eventual equilibrium is likely to be beyond anybody's vision, including that of the most enlightened regulator.) The danger of these regulatory interferences is greatest where these actions foreclose competitive options during the transition. Another suggestion, actually embodied in the *Staggers Act*, is to move away from the prior regulated structure only slowly, and then in a rigidly specified fashion, wherever substantial segments of captive consumers continue to exist. Still another thought is to keep particular rates as much in line as possible with the underlying long-run marginal costs of supplying a particular activity, despite the possible short-run welfare advantages of certain price discrimination schemes. It should never be forgotten that in a market economy, whenever prices for any activity rise disproportionately above the underlying cost fundamentals, business ingenuity (in the form of new or altered technologies, product and location substitutions, etc.) is quickly applied to finding ways to do with less. In many ways, this is the fundamental lesson of the long and tortured history of ICC regulation of the railroads, to a lesser

extent of FCC regulations of long-lines telecommunications, and most recently of the OPEC cartel's efforts to set international oil prices. To paraphrase Keynes—in the long run not only are we all dead but with provision of enough incentives, all demands are elastic.

## REFERENCES

- Baumol, William J. and Bradford, David J., "Optimal Departures from Marginal Cost Pricing," *American Economic Review*, December 1970, 60, 265–83.
- Baumol, William J. and Willig, Robert D., *Ex Parte* No. 347 (Sub-No. 1), *Coal Rate Guidelines—Nationwide*, before the ICC, July 28, 1983.
- \_\_\_\_\_, \_\_\_\_\_, and Goldfeld, Stephen, "Verified Statements" (on behalf of the Eastern railroads), May 11, 1981; Kenneth Arrow, Leon N. Moses, Ronald R. Braeutigam, and William Wecker, "Verified Statements" (on behalf of the Western railroads), May 11, 1981; in *Ex Parte* No. 347 (Sub-No. 1), *Coal Rate Guidelines—Nationwide*, before the ICC.
- Baumol et al., William J., "The Role of Cost in the Minimum Pricing of Railroad Services," *Journal of Business*, July 1962, 36, 348–51.
- Gordon, Kenneth, "Deregulation, Rights, and the Compensation of Losers," in Kenneth D. Boyer and William G. Shepherd, eds., *Economic Regulation: Essays in Honor of James R. Nelson*, East Lansing: Michigan State University Press, 1981.
- Hotelling, Harold, "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates," *Econometrica*, July 1938, 6, 242–69.
- Kahn, Alfred E., "Applications of Economics to an Imperfect World," *American Economic Review Proceedings*, May 1979, 69, 1–13.
- Levin, Mark M. and Stram, Bruce N., "Nursing the Railroads Back to Health," *Regulation*, September/October 1981, 29–36.
- Locklin, D. Phillip, "The Literature on Railway Rate Theory," *Quarterly Journal of Economics*, February 1933, 47, 167–230.
- Meyer, John R., "Toward a Better Understanding of Deregulation: Some Hypotheses and Observations," *International Journal of Transport Economics*, April/August 1983, 10, 39–41.
- Meyer et al., John R., *The Economics of Competition in the Transportation Industries*, Cambridge: Harvard University Press, 1959.
- Panzar, John C. and Willig, Robert D., "Economies of Scale in Multi-Output Production," *Quarterly Journal of Economics*, August 1977, 91, 481–93.
- \_\_\_\_\_, and \_\_\_\_\_, "Economies of Scope," *American Economic Review Proceedings*, May 1981, 71, 268–72.
- Ramsey, Frank P., "A Contribution to the Theory of Taxation," *Economic Journal*, March 1927, 37, 47–61.
- Tye, William B., (1983a) "Balancing the Ratemaking Goals of the Staggers Rail Act," *Transportation Journal*, Summer 1983, 22, 17–26.
- \_\_\_\_\_, (1983b) "Ramsey Pricing and Market Dominance Under the Staggers Rail Act of 1980," *Transportation Research Forum, Proceedings—Twenty-Fourth Annual Meeting*, 1983, 667–74.
- \_\_\_\_\_, "The Role of Revenue/Variable Cost Ratios: Rail Market Dominance Determinations," *Transportation Journal*, Winter 1984.
- \_\_\_\_\_, "Problems of Applying Stand-Alone Costs as an Indicator of Market Dominance and Rate Reasonableness," *International Journal of Transport Economics*, February 1985.
- \_\_\_\_\_, and Leonard, Herman, "On the Problems of Applying Ramsey Pricing to the Railroad Industry with Uncertain Demand Elasticities," *Transportation Research*, November 1983, 17A, 439–50.
- Willig, Robert D., "Verified Statement," Docket No. 37850S, before the ICC, June 3, 1983.
- ICC, Unpublished *Decision*, decided February 8, 1983, *Ex Parte* No. 347 (Sub-No. 1), *Coal Rate Guidelines—Nationwide*.

# "Let Them Make Toll Calls": A State Regulator's Lament

By ROGER G. NOLL\*

In the late 1960's, the Federal Communications Commission (FCC) introduced competition into telecommunications. Initially limited to specific services and types of customer equipment, the limits soon gave way. By 1980, the FCC's policy was to promote competition.

In 1982, the Antitrust Division settled its suit against AT&T with close to total victory, achieving divestiture of the Bell Operating Companies (BOC). AT&T remains in equipment and interexchange services, which are growing increasingly competitive. To facilitate divestiture, the FCC adopted several policies: asserting jurisdiction regarding depreciation and then adopting methods that more nearly reflect economic costs, eliminating regulation of equipment prices, and restructuring the procedures whereby interstate services share local exchange costs.

Two aspects of these new policies are worth emphasizing. First, astonishingly enough, economics played a central role in changing federal telecommunications policy, as acknowledged by Philip Verveer (1984), the lawyer who developed the antitrust case against AT&T, the Chief of the FCC's Cable Television Bureau when cable was deregulated, and the Chief of the Common Carrier Bureau when the FCC formally adopted the policy of minimizing federal regulation of telecommunications. The intellectual foundation of these policies is an economic case that the industry will be more efficient if it is minimally regulated and maximally competitive.

Second, the new federal policy is widely despised by state regulators. My title is from an eloquent decision in Texas, which also characterized cost-causative pricing as from the "Marie Antoinette School of Rate Design" (Mary Ross McDonald and Angela

Marie Demerle, 1984, p. 35). State regulators dislike federal procompetitive policy because it transferred several billion dollars of revenue responsibility to the states and threatens state regulatory policies. Thus far, the state response has hardly been accommodative. Instead, federal and state regulators are fighting a three-front "Jurisdiction War." This paper briefly analyzes the economics and politics of state resistance to federal policies. For more details, see my companion paper (1985).

## I. The Economics of State Telecommunications Regulation

The source of the conflict lies in the methods that states use to regulate telecommunications. The most important features of state ratemaking are allowable cost estimation, interjurisdictional revenue sharing, rate averaging, and residual pricing.

*Allowable Cost Estimation.* Regulators are supposed to assure that regulated service is provided at reasonable cost and that a regulated utility recovers useful investments and earns reasonable profits. Regulation is not public enterprise; firms have considerable latitude in making normal business decisions. Commissions normally ascertain whether investments are needed, and sometimes determine whether the prices paid by utilities are reasonable, but they rarely question decisions regarding network design or the technical characteristics of services. Once the investment is approved, its costs and profit enter permanently into the firm's revenue requirement.

The key consequences of allowable cost practice are that it is unlikely to force a utility to minimize the cost of service, and that investment risks are not borne by the utility. Hence, the allowed cost of an investment need bear no relation to its economic value. Should unanticipated changes in technology or market structure render an asset

\*Stanford University, Stanford, CA 94305.

uneconomic, it nonetheless remains in the rate base, and regulators are obligated to find sources of revenue for it.

Divestiture and federal deregulation apparently have left local telephone companies with substantial uneconomic, incompletely amortized assets. Virtually all equipment owned by BOCs was manufactured by Western Electric. Judging from postdivestiture events, Western apparently was a high-cost supplier of several important lines of equipment. Since divestiture, Western has lost market share, closed manufacturing facilities, and engaged in cost-reducing actions to compete for business. If the BOCs paid super-competitive prices for equipment, their books overstate the economic value of their assets. Accounting practices in telecommunications also work to overstate asset values. Examples are excessively slow rates of depreciation and capitalization of some recurring expenses (for details, see Nina Cornell and myself, 1985).

The problem for state regulators is that competition severely constrains, if not undermines, the prospects of paying for uneconomic assets. As the FCC deregulates interstate toll, customer equipment, cable television, domestic satellites, and several technical challenges to the local exchange system, the state's ability to maintain regulated intrastate prices above economic cost diminishes.

*Interjurisdictional Revenue Sharing.* Most interstate telecommunications service connects to the interstate network through the local exchange. Regulators have developed formulas whereby interstate services pay for this use. The procedure divides local exchange costs into those that vary with use and those that are nontraffic sensitive (*NTS*). The former are allocated among jurisdictions on the basis of relative minutes of use (for example, fully allocated costs). In the past, the latter were allocated according to an economically meaningless formula which caused interstate service to pay about 25 percent of local *NTS* costs with this fraction growing rapidly. The formula was negotiated among the FCC, state regulators, and the industry, but the FCC unilaterally terminated it, capped the federal contribution at 25 percent, and changed the method of raising the

interstate *NTS* contribution. The FCC had imposed an excise tax on interstate toll, based on minutes of use. In 1984, the FCC adopted an "access charge"—a fixed monthly payment for access to long distance through the local *NTS* plant.

Historically, states have regulated the monthly rate for local service. The FCC's access charge adds between \$2 and \$6 to this rate. Economically, although the issue is complicated (see Robert Willig, 1979), a hookup charge is a sensible way to cover *NTS* costs; politically, state regulators see it as transferring federal *NTS* charges to the states. Regulators believe that customers are unlikely to grasp the subtle distinction between the federal and state components of basic rates. Functionally, the FCC's decision uses the monthly flat rate to cover the costs of uneconomic assets. Thus, as competition made covering the state share of these costs more difficult, the FCC preempted one possible escape by raising the flat monthly rate.

*Rate Averaging.* Common regulatory practice sets the same price for any given service for all customers. Consequently, basic service rates do not reflect differences in the cost of service.

Although *NTS* costs vary for many reasons, the most important cause is population density. Most *NTS* costs are copper wires connecting a customer to the first switch in the network. In sparsely populated areas, connection distances can be several times as great as in urban areas. Thus, rate averaging creates a massive subsidy of rural service. The average federal contribution to *NTS* costs for business and residential users is about \$6 per month, but in rural states it is much greater, reaching \$25 per month in Wyoming (Congressional Budget Office, 1984).

Some investment in rural service probably is uneconomic. Rural residents may be unwilling to pay for telephone service that is priced at its accounting cost. Moreover, copper-wire technology probably is not the cheapest way to serve rural areas. Instead, recent technical advances probably make over-the-air technologies, such as cellular radio, cheaper in some areas than the book value of a rural *NTS* plant. If so, the eco-

nomic value of an existing plant falls short of its book value.

The regulator's problem is to find revenue to pay for expensive rural service. The FCC's access charge places a greater burden of cross subsidization on the monthly flat rate, with nontrivial distributional consequences across categories of users. Moreover, it underscores the interstate subsidy flow. State regulators in the urbanized Northeast believe that residences and small businesses in their jurisdiction will pay a larger, more visible subsidy to western ranchers.

*Residual Pricing.* The plight of the states is largely of their own doing, resulting from traditional ratemaking philosophy: residual pricing. Economics, efficiency, and cost causation play a tiny role in intrastate rate-making. Instead, pricing begins by designating "basic" services as having special social importance. The installation charge and the monthly flat rate for local service are accorded such status everywhere, and in some states so is the price of local pay calls. Within the limits of economic feasibility (for example, demand elasticities) and "fairness," an increase in revenue requirements is covered by price increases for nonbasic services. The remaining revenue requirement—the residual—is covered by price increases in basic services.

Because residual pricing considers only total costs, the prices of nonbasic services are set roughly to maximize gross revenues, subject to a loose political constraint on the acceptable magnitude of a price increase. This implies demand elasticities for nonbasic services at or near unity, and much lower demand elasticities for basic service. Residual pricing also permits pricing some nonbasic services below economic costs. Because services do not face a meaningful cost test, the price that maximizes gross revenue need not be higher than average or marginal cost. Again, although data are fragmentary, evidently some intrastate services, notably private lines and Centrex, are so priced.

The argument supporting residual pricing has been the "universal service" objective, where this means maximizing the number of subscribers to telephone service. State regulators fear that cost-causative prices for basic

service would cause significant disconnections from the telephone network. This fear is not justified. Briefly, estimates of the demand elasticity for basic service are very low, and prices based on economic costs probably would not differ much from current prices, except in rural areas.

Residual pricing worsens the problems facing state regulators. As they seek to cover revenue requirements, they begin with prices for most services that roughly maximize gross revenues. Moreover, competition threatens to make firm-specific demand curves more elastic. Only the politically most visible service has a small demand elasticity, and here, the FCC has preempted the first big price increase.

## II. Jurisdiction Wars

Policy conflicts between state and federal regulators take many forms: equity vs. efficiency, regulation vs. competition, residual pricing vs. cost-causative pricing, etc. Underlying these is the deeper problem of paying for undepreciated uneconomic assets and expensive rural service when technological change and competition are increasingly likely to attack overpriced services.

The process by which these conflicts are revealed and resolved is a jurisdiction war. The condition necessary for a jurisdiction war is that two regulatory authorities have conflicting policies, and sufficiently overlapping responsibilities that decisions by one substantially affect the ability of the other to achieve its objectives.

For example, the FCC seeks a competitive interexchange toll industry. The companies providing interstate service can use their facilities to compete for local toll. If states handicap local competition to retain a source of revenues for local exchange companies, they inhibit FCC policy and raise the cost of toll in the FCC's jurisdiction.

The FCC also promotes competition in customer equipment. One example is the PBX, a switch on a customer's premises that connects office extensions. Customer PBX competes with Centrex, a service offered by local exchange companies that uses the local network for interextension calls. If state reg-



ulators underprice Centrex, they destroy PBX competition; however, if state regulators price Centrex above costs, PBX sales erode its market. This situation illustrates the efficiency benefits of competition, but also demonstrates how decisions by one jurisdiction affect the other, and hence create inter-jurisdictional conflict.

Jurisdiction wars are fought on three fronts. First, one agency adopts regulations that undermine the other's policies. For example, low federal prices for interstate private line service constrain intrastate toll prices. States cannot prevent the use of interstate private lines and return interstate toll to reach an intrastate destination.

Second, one agency asserts jurisdiction over another or in an unregulated area. Examples are depreciation, where the FCC took responsibility away from the states, and cellular radio, where the FCC used its authority to allocate the electromagnetic spectrum to introduce two-firm rivalry, whereas states would prefer to make cellular radio part of the local telephone monopoly.

Third, one agency attacks another through political overseers, especially legislators. In 1984, after the access charge decision, state regulators lobbied Congress for relief. Congress convinced the FCC to delay its decision for residential users until early 1985. Previously, state regulators unsuccessfully sought legislation to limit deregulation and to undermine the antitrust cases against AT&T.

For economic, technical and political reasons, the states appear destined to lose the jurisdiction war on all three fronts. In interactive regulations, the FCC has economics and technology on its side. Competition and, where there is regulation, cost-causative pricing will undermine state policies wherever the two overlap, for the former will cause local exchange companies to retain subsidized activities while losing subsidizing business. Moreover, by regulating the electromagnetic spectrum and cable television, the FCC controls the technologies that most threaten the local telephone network.

The FCC also holds the cards in direct battles over jurisdiction. The courts permit the FCC to assert jurisdiction wherever it can show a connection between its statutory

policy responsibilities and the domain it seeks to regulate. The FCC acquired jurisdiction over cable television because cable affected the regulation of broadcasting, and over depreciation of the local exchange because it affected interstate pricing. Most likely, the FCC could assert jurisdiction over all of telecommunications, for the interconnectedness of the network makes impossible a clean, noninteractive jurisdictional separation.

On the political front, the states have the best hope for success. Mayors and state office holders are the potential competition for incumbents in congress. If local telephone rates become politically salient, the no-risk strategy for Congress is to preempt the issue by restoring the status quo ante.

Two factors make this unlikely. First, rate increases are unlikely to be large enough to become one of the few issues that are salient in legislative elections. Second, the politics of telephone pricing is likely to turn against most state regulators as data reveal that most of the problem is traceable to a rural subsidy.

Starting with *Baker v. Carr* in 1962, the Supreme Court issued several decisions clarifying the constitutional requirements for legislative districts at all levels of government. The enunciated policy, "one man, one vote" in that gender insensitive era, massively rearranged legislative districts and vastly reduced the legislative influence of rural areas (see Mathew McCubbins and Thomas Schwartz, 1984). It also caused reductions in rural subsidies, or a transformation of them to programs with a broader base (for example, food stamps).

The institutionalization of pricing policies that subsidize rural telephone service took place before redistricting. Moreover, rapidly advancing technology caused real telephone prices to fall throughout the 1970's, thereby attracting no political attention. Following Paul Joskow's (1974) observations about the asymmetry of political responses to increases vs. decreases in regulated prices, new federal policy focuses political attention on the intrastate prices that will increase. In the information-poor world of telecommunications regulation where the incidence of costs and subsidy burdens is obscure, the initial politi-

cal response is to view these increases with alarm. As political leaders receive information that the issue is yet another rural subsidy, the politics should change. This does not imply that rural subsidies will end, but that political pressure will emerge to develop a more targeted, smaller rural subsidy, perhaps with help from legislative appropriations.

The final interesting question is whether state regulators will move quietly toward less regulation and more efficient prices, or will go down fighting. The longer state regulators delay, the more costly will be the change when it comes because with competition, cross subsidization will cause inefficient patterns of investment by entrants and customers. Politics is normally reactive, so rational state policy probably requires that regulators be in front of political leaders, and take early actions that displease them. Such bureaucratic entrepreneurship has a precedent, for it was practiced by several federal regulatory agencies during the 1970's, including the FCC.

#### REFERENCES

- Cornell, Nina W. and Noll, Roger G., "Local Telephone Prices and the Subsidy Question," Stanford Studies in Industry Economics, January 1985.
- Joskow, Paul, "Inflation and Environment Concern," *Journal of Law and Economics*, October 1974, 17, 291-327.
- McCubbins, Mathew and Schwartz, Thomas, "The Politics of Derustication," Stanford Center for Economic Policy Research Publication No. 30, August 1984.
- McDonald, Mary Ross, and Demerle, Angela Marie, "Examiners' Report: Docket No. 5113," Public Utility Commission of Texas, April 2, 1984.
- Noll, Roger G., "State Regulatory Responses to Competition and Divestiture in the Telecommunications Industry," in Ronald E. Grieson, ed., *Antitrust and Regulation*, Lexington: Lexington Books, 1985.
- Verveer, Phillip, "Regulation and the Access Problem: What's Happened and Where We Are Now," in Alan Baughcum and Gerald Faulhaber, eds., *Telecommunications Access and Public Policy*, Norwood: Ablex, 1984.
- Willig, Robert D., "The Theory of Network Access Pricing," in Harry M. Trebing, ed., *Issues in Public Utility Regulation*, East Lansing: Michigan State University, 1979.
- Congressional Budget Office, *The Changing Telephone Industry: Access Charges, Universal Service, and Local Rates*, Washington: USGPO, June 1984.

# Intercity Transportation Route Structures under Deregulation: Some Assessments Motivated by the Airline Experience

By STEVEN A. MORRISON AND CLIFFORD WINSTON\*

During the past decade, the United States intercity transportation system has undergone significant changes owing to legislation that effectively deregulated the major transportation industries. This paper assesses the effects of transportation deregulation on carriers' route structures, using as motivation experience from airline deregulation. Route structure changes have important implications for the level of service that many travelers and shippers will receive in the new transportation environment. We first summarize the major results from our (1984) study on the welfare effects of airline deregulation. The implications of these findings are then used as a basis for analyzing the route structures of the major modes in the deregulated environment.

## I. Welfare Effects of Airline Deregulation

Our earlier paper was concerned primarily with analyzing the welfare effects on passengers of airline deregulation. We assembled a sample of passenger trips over roughly 800 different routes, encompassing every level of passenger flow density based on hub classification.<sup>1</sup> Utilizing this data base, we calcu-

lated the change in business and pleasure travelers' welfare (i.e., compensating variations) that resulted from fare, entry, and exit deregulation. Our major finding is that, in 1977 dollars, the welfare increase for all travelers is \$5.7 billion. Decomposing this net welfare change, we find the average welfare change per passenger trip due to fare changes is \$4.04, to travel time changes is -\$0.96, and to frequency changes is \$8.00.

Although our finding of a substantial welfare gain from deregulation is consistent with economists' policy recommendations and predictions (see Winston, 1985), the explanation that underlies it is not. Conventional wisdom (pioneered by George Douglas and James Miller, 1974) held that regulation led to excessive fares accompanied by superior service quality, particularly departure frequency, compared to what would be generated in a deregulated environment. Deregulation was therefore believed to be socially beneficial because it would lead to fare reductions that would more than compensate for reductions in service quality. We found, however, that the largest source of the welfare gain from deregulation occurred (on average) through increases in departure frequencies.<sup>2</sup>

A summary of our data, presented in Table 1, shows the improvements in departure frequencies that resulted from deregulation. These percentage changes, which include direct and connecting flights, correspond to a route-weighted average increase in frequen-

\*Department of Economics, Northeastern University, Boston, MA 02115, and Economic Studies Program, The Brookings Institution, 1775 Massachusetts Avenue, NW, Washington, D.C. 20036, respectively. We are grateful to Joan Winston for significant contributions to this research.

<sup>1</sup>An individual community falls into one of four hub classifications based on that community's percentage of total enplaned revenue passengers at U.S. airports. Those communities enplaning 1 percent or more of the total are classified as large hubs, communities enplaning between .25 and .99 percent of the total are classified as medium hubs, communities enplaning between .05 and .24 percent of the total are classified as small hubs, and communities enplaning less than .05 percent of the total are classified as nonhubs. In 1977, there were 25 large hubs, 39 medium hubs, 94 small hubs, and 471 nonhubs.

<sup>2</sup> This finding assumes that pleasure travelers paid the median discount fare under both regulation and deregulation, and that all business travelers paid the coach fare under regulation, but one-half paid the coach fare and one-half paid the median discount fare under deregulation. The assumed proportions of discount fare travel are consistent with Air Transport Association *Monthly Discount Reports*.

TABLE 1—WEIGHTED AVERAGE PERCENTAGE CHANGE  
IN FLIGHT FREQUENCY BY HUB PAIR CLASSIFICATION  
1977-83

	Nonhub	Small Hub	Medium Hub	Large Hub
Nonhub	33.9	1.4	24.3	28.7
Small Hub		33.9	20.8	19.2
Medium Hub			-4.3	14.4
Large Hub				-3.5

cies of 9.2 percent. By way of comparison, aircraft departures increased by 1.5 percent. The increase in convenient connecting flights, leading to additional departure alternatives, is a result of a major change in airline operations. Under deregulation, airlines have adopted hub-and-spoke route structures, where travelers are fed via spoke routes into a major airport (hub) from which they proceed to take connecting flights to their destinations. Hub-and-spoke route structures were not developed fully under regulation because of entry restrictions and insufficient incentives, as perceived by management, to generate profit by a major restructuring of airline networks. Our results suggest that this change in operations contributed greatly to the success of deregulation. Because it is likely that the adoption of hub-and-spoke route structures is one of the major consequences of airline deregulation, it is of interest to analyze under what conditions, and for what modes, this type of route structure or alternative route structures is desirable.

## II. Intercity Transportation Route Structures

We illustrate our analysis of the conditions that favor hub-and-spoke route structures with reference to Figure 1, which represents a transportation carrier operating over a simple network. By regulatory authority, the carrier serves route 1 ( $A-C$ ) supplying output  $Y_1$ , (measured in passengers or tons of freight) and serves route 2 ( $B-C$ ) providing output  $Y_2$ . For simplicity, there is assumed to be no traffic on or regulatory authority to operate between the  $A-B$  segment. Under deregulation, the carrier faces three alternatives regarding the choice of optimal route structure that are of interest here. The carrier can

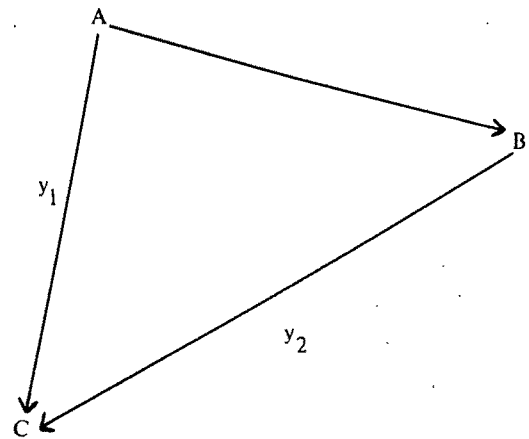


FIGURE 1

abandon service on one route (say route 1), it can maintain the same route structure as under regulation (status quo), or it can adopt a hub-and-spoke structure where traffic originating from  $A$  destined for  $C$  is routed through  $B$  and combined with traffic originating at  $B$  destined for  $C$ . The profit equations,  $\Pi^i$ , that correspond to these alternatives are

- (1)  $\Pi^1 = P_2 Y_2 - C(0, Y_2)$
- (2)  $\Pi^2 = P_1 Y_1 + P_2 Y_2 - C(Y_1, 0) - C(0, Y_2)$
- (3)  $\Pi^3 = P_1 Y_1^* + P_2 Y_2 - C(Y_1^*, Y_2)$ ,

with (1) being abandonment, (2) status quo, and (3) hub-and-spoke, and where  $P_i$  denotes price in market  $i$  (assumed constant across alternatives),  $C(\cdot)$  denotes total cost,  $Y_1^*$  denotes output from market 1 when routing traffic through the hub, and  $Y_1^* \leq Y_1$  because hub-and-spoke practices will increase travel time for traffic originating at  $A$  and thus possibly reduce output.<sup>3</sup>

We first consider the optimal route structure assuming two markets are to be served,

<sup>3</sup> In a more complicated network, traffic consolidation will increase the frequency of direct or connecting service relative to nonstop (point to point) service, thus possibly offsetting the effect of increased travel time.

and then examine under what conditions a carrier would abandon service. Given both markets are to be served, the criterion for adopting a hub-and-spoke route structure based on (2) and (3) is whether

$$(4) \quad C(Y_1, 0) + C(0, Y_2) - C(Y_1^*, Y_2) \\ > P_1(Y_1 - Y_1^*).$$

A hub-and-spoke route structure is thus desirable if the cost saving from producing output 1 jointly with output 2 (as opposed to producing output 1 independently) exceeds the possible loss in revenue.

A more complete understanding of this criterion can be obtained by considering the key determinants of  $C(Y_1^*, Y_2)$ . These joint costs will be reduced (increased) relative to independent production if the cost savings from economies of aircraft (or vehicle) size are greater (lower) than the cost of rerouting traffic. (In addition to the cost of increased distance traveled, more elaborate passenger or freight handling facilities will be required at the hubs.) That is, whether

$$(5) \quad C(Y_1^*|BC) + C(Y_2|BC) - C(Y_1^*, Y_2|BC) \\ > C(Y_1^*|AB + BC) - C(Y_1^*|AC),$$

where  $C(Y_1^*|BC)$  denotes the cost of transporting  $Y_1^*$  passengers (freight) on the route  $B$  to  $C$  and the other cost expressions are interpreted accordingly. The left-hand side of the inequality denotes vehicle size economies and the right-hand side denotes the cost of rerouting traffic.

To summarize, adoption of a hub-and-spoke route structure is warranted if scope economies are sufficient to offset any losses in revenue. The existence and magnitude of scope economies is a function of economies of aircraft (or vehicle) size and rerouting costs, while revenue loss depends on the resulting change in travel time (and frequency) and the users' relevant demand elasticities.

Generally, all modes are characterized by economies of vehicle size. (For discussions of these economies see Douglas and Miller for aircraft, Motor Vehicle Manufacturers Association, 1983, for buses and trucks, and

Theodore Keeler, 1983, for railroads.) These economies derive from more efficient use of labor and fuel associated with vehicles with greater capacity. Because of particular labor practices and vehicle dynamics, these economies are slightly greater for aircraft and trains, although buses and trucks exhibit significant vehicle size economies. For example, the Motor Vehicle Manufacturers Association reports that a 20 percent increase in productivity can be achieved by using double-trailer trucks instead of single-trailer trucks, while similar productivity gains can be achieved by using articulated buses instead of standard buses.

As a first approximation, rerouting costs are proportional to increases in travel distance for all modes. Thus, given that the potential for cost savings from hub-and-spoke operations appears to be similar across modes, the *relative* attractiveness of these operations depends largely on their effect on each modes' demand (i.e., revenue).

The final issue concerns whether a carrier should abandon service. Generally, a carrier should abandon service in a market if the fixed cost of operations exceeds the difference between the marginal revenue and corresponding marginal cost for that market. As illustrated in Richard Ericson and Winston (1983), the factors that can encourage a carrier to exit a market are, on the supply side, the presence of strong competition and rising marginal costs, and on the demand side, a high-market demand elasticity and a low price or weak demand.

It is now possible to assess or predict the impact of deregulation on various modes' route structures. Given air transportation technology, the change to hub-and-spoke operations is understandable: the economies of aircraft size that contribute to scope economies generally are not offset by revenue losses, because increases in travel time and travelers' corresponding responses are small (see our 1985 article). It is also interesting to note that the nature of hub-and-spoke operations can discourage abandonments because the benefits from these operations are derived from establishing and maintaining spoke routes to realize economies of aircraft size. Service is also preserved because major

airlines frequently coordinate with commuter airlines to serve cities with insufficient traffic to warrant service by the large aircraft used by major carriers.

The effect of the 1983 intercity bus deregulation legislation on route structure is likely to differ from the air experience. Bus technology does not appear to encourage the adoption of hub-and-spoke practices because any significant rerouting will increase travel time by a large amount and thus lead to a substantial reduction in patronage (see our 1985 article). This suggests that bus carriers are likely to abandon some routes under deregulation. That is, the most appropriate bus route structure is characterized by service of cities along a linear network. Some routes will not be economically justified for service and are thus likely to be abandoned.

Motor freight transportation was largely deregulated in 1980. Because deregulation eliminated vehicle routing restrictions, a hub-and-spoke type route structure has evolved significantly for less-than-truckload (*LTL*) service. Operationally, *LTL* shipments from several originating points are consolidated at major terminals (hubs) and sent out to their destinations. This practice is appropriate for motor freight transportation because economies of vehicle size are significant, while the cost of re-routing shipments is not excessive. Moreover, the increase in shipment transit time from this practice is not likely to reduce traffic substantially (see Winston, 1981). The adoption of hub-and-spoke type operations for *LTL* service has also preserved service on many routes.

Potential changes in the rail passenger transportation environment suggest our analysis may be useful in predicting future rail route structures. Currently, there is serious consideration of high-speed rail transportation in many parts of the country. In addition, latent concerns about rail passenger subsidies may eventually resurface and lead to the transfer of responsibility for regular rail service from the public to the private sector. In light of these possibilities, our analysis suggests that given the nature of rail technology, particularly the constraints imposed by the necessity of a fixed right-of-way, high-speed, and regular rail routes will corre-

spond to a linear network. Current routes that are not economically justified in such a network are likely to be abandoned. Rail route abandonments could be minimized if a modified hub-and-spoke route structure were developed, where bus transportation serviced passengers on spokes in combination with rail service on the main linehaul.

Finally, one of the major outcomes of the 1980 rail freight deregulation legislation has been the abandonment of service on many routes. As in passenger service, rail freight service is best carried out primarily for a linear network. It is possible, however, that many abandoned routes could eventually receive rail service through the development of coordinated rail-motor carrier hub-and-spoke type operations where motor carriers connect these routes with rail linehaul service.

The recent flood of transportation deregulation legislation will ultimately have different effects on modes' route structures. The benefits from deregulation depend significantly on how regulated route structures differ from optimal route structures. Some modes (air, motor carrier) have adopted hub-and-spoke operations that improve, or at least maintain, service to outlying (or low density) routes, while other modes (rail, bus) will develop their linear networks and abandon economically unjustified service to such routes. As a result, many shippers and travelers on outlying routes may have to depend on the development of joint modal operations to avoid a partial or total loss of their pre-deregulation transportation services. The actual value of such losses, should they occur, and their effect in assessments of the welfare effects of surface freight and passenger deregulation, await further research.

## REFERENCES

- Douglas, George W. and Miller, James C. III, *Economic Regulation of Domestic Air Transport*, Washington: The Brookings Institution, 1974.
- Ericson, Richard and Winston, Clifford, "Predatory Capacity Expansion in a Deregulated Motor Carrier Industry," *Research in Transportation Economics*, 1983, 1, 185-235.

Keeler, Theodore E., *Railroads, Freight, and Public Policy*, Washington: The Brookings Institution, 1983.

Morrison, Steven A. and Winston, Clifford, "An Econometric Analysis of the Demand for Intercity Passenger Transportation," *Research in Transportation Economics*, forthcoming, 1985.

\_\_\_\_\_ and \_\_\_\_\_, "The Welfare Effects on Travelers of Airline Deregulation," Working Paper, 1984.

Winston, Clifford, "A Disaggregate Model of the Demand for Intercity Freight Transportation," *Econometrica*, July 1981, 49, 981-1006.

\_\_\_\_\_, "Conceptual Developments in the Economics of Transportation: An Interpretive Survey," *Journal of Economic Literature*, March 1985, 23, 57-94.

Motor Vehicle Manufacturers Association, *Building The Tools that Move America*, Detroit: MVMA, 1983.

## THE END OF THE GREAT BOOM AND THE BREAKDOWN OF BRETTON WOODS: WAS IT A COINCIDENCE?<sup>†</sup>

### Macroeconomic Stability and Flexible Exchange Rates

By JOHN F. O. BILSON\*

During the past decade, the monetary approach to the theory of exchange rate determination has evolved rapidly in response to new empirical findings concerning the behavior of international financial markets. In addition to the founding assumptions of a stable money demand function and an exogenous money supply, the early monetary models relied upon three arbitrage conditions as market-clearing equations. The purchasing power parity condition integrated national commodity markets; the interest rate parity condition integrated national bond markets; and the forward parity condition provided an essential link between spot and forward markets by stating that the forward price was the market estimate of the future spot price. These arbitrage conditions distilled the complexity of a full general equilibrium model into a single equation which stated that the spot exchange rate was determined by the expected future path of the relative ratio of money to income.<sup>1</sup>

The purchasing power parity condition was the first victim of the empirical realities of the post-Bretton Woods system. Because of the magnitude of the variation in real exchange rates, and because the variation in exchange rates vastly exceeded the variation in relative commodity prices, it became obvious that the exchange rate did not clear the market by its indirect influence on relative prices and commodity flows. The failure of the price arbitrage condition was an im-

portant contributor to the development of the "asset market" approach in which financial markets were assumed to be "efficient" while commodity markets were characterized by sluggish price adjustment. The most widely used model of this type is, of course, Rudiger Dornbusch's model (1976) of exchange rate dynamics.

From the perspective of this paper, the most important result that Dornbusch derived was the "overshooting effect," in which sluggish price adjustment in the commodity (and labor) markets resulted in increased variability in exchange rates and interest rates. The result is important because it demonstrated that the instability of the floating rate system could be due to the inherent differences between commodity and financial markets. This view stands in strong contrast to the libertarian consensus which attributes all of the instability of the system to the irrational practices of the public participants in the market.

Both the asset market and monetary approaches relied upon the forward parity condition as the link between current and expected future prices. If the forward price could be assumed to be a market forecast of the future spot price, then the techniques developed in the rational expectations literature could be used to find concise expressions for the relevant endogenous variables. For those with an empirical bent, the forward parity condition became the next target after the purchasing power parity theory had been beaten to death.

Since it is impossible to exactly determine the expected future spot rate, it is empirically more difficult to reject the forward parity condition. However, as the number of observations has increased and as econometric techniques have become more sophisticated,

<sup>†</sup>*Discussants:* Jo Anna Gray, Washington State University; Paul Krugman, Massachusetts Institute of Technology.

\*Graduate School of Business, University of Chicago, Chicago, IL 60637.

<sup>1</sup>For a review of these models, see my article (1979).



the validity of forward parity has become increasingly questionable. The forward premium, the logarithmic difference between the forward and spot exchange rates, can be tautologically divided into a risk premium and an unbiased forecast of the rate of change in the exchange rate. The forward parity condition assumes that the major part of the variation in the premium is due to the forecast element, and that international risk premia are small and stable over time. The new empirical evidence rejects this view and replaces it with the empirical paradigm that most of the variation in the premium reflects variation in the risk premium rather than variation in the expected rate of appreciation.<sup>2</sup>

This finding shall form the basis of this paper's analysis of the instability of financial prices under floating rates. Specifically, the object of the paper is to determine if the failure of the forward parity condition is a cause of instability in the same way that the failure of purchasing power parity is a cause of instability in the Dornbusch model. In order to investigate this issue, it is necessary to first define the failure of forward parity in a more specific manner.

Consider the following regression equation:

$$(1) \quad s_{t+1} - s_t = \beta(f_t - s_t) + u_t,$$

where  $s(f)$  represents the log of the spot (forward) exchange rate,  $\beta$  is a regression coefficient,  $u$  is the residual, and  $t$  is an index of time which, for this discussion, shall be assumed to be measured in units of one month. It is also assumed that the residuals are identically and independently distributed. Equation (1) can be considered as a composite forecast of the future spot rate. The composite includes the current spot rate, which would be the appropriate forecast if the exchange rate evolved as a random walk, and the forward rate, which would be ap-

propriate if the forward parity model was correct. Within this interpretation, the regression coefficient may be interpreted as the weight on the forward rate in the composite forecast.

Recently, Eugene Fama has suggested a different interpretation of this equation. Using the decomposition of the forward premium into its risk premium and expectational components, Fama demonstrates that the regression coefficient may be used to determine the relative variability of the two components. If the estimate is unity, then all of the variation in the premium may be attributed to expectations; if it is zero, then all of the variation in the premium can be attributed to variation in international risk premia.

The recent empirical evidence has led to a reasonably firm rejection of the hypothesis that the true value of the  $\beta$  coefficient is unity. Based upon my own subjective review of the literature, my prior mean for  $\beta$  is about .2 with 95 percent confidence limits ranging from -.3 to .7. It is consequently not possible to reject the view that the forward premium, and hence international differences in short-term interest rates, are unrelated to either actual or market forecasts of exchange rates. The next major task for exchange rate theorists, then, is to develop an alternative model of the determination of the premium and to incorporate this alternative into the existing literature on exchange rates.

My purpose in this presentation is more limited. The question that I wish to address is the following: if the forward parity assumption was correct, would spot exchange rates and forward premia be more or less stable? The issue can also be stated negatively: has the failure of forward parity been an independent cause of financial volatility? My starting point is the standard asset market reduced-form equation:

$$(2) \quad s_t = z_t + \phi(f_t - s_t),$$

where  $z_t$  represents the current fundamental influences on the exchange rate. (In the monetary model,  $z_t$  is equal to the relative money to income ratio of the two countries.) If the exchange rates are quoted in dollar

<sup>2</sup>My paper with David Hsieh (1984) and that by Eugene Fama (1984) present estimates of the equation in the text. There are many papers with estimates of similar equations.

terms,  $f_t - s_t$  is the forward premium on the currency from a dollar perspective. Since a positive forward premium implies a lower local interest rate and hence a lower opportunity cost of holding the currency, it is reasonable to assume that the  $\phi$  parameter is positive. If we assume that  $\phi$  is equal to Cagan's semi-elasticity, and that the interest elasticity of the demand for money is .15 when nominal interest rates are 1 percent per month, then the value of  $\phi$  should be around 15. This estimate will prove useful in the discussion below.

Combining equation (2) with expectations of the future spot rate formed on the basis of equation (1), the spot rate can be expressed as a combination of the current fundamentals,  $z_t$ , and the expected future spot rate,  $E_t s_{t+1}$ :

$$(3) \quad s_t = (1 - \gamma)z_t + \gamma E_t s_{t+1},$$

where  $\gamma$  is defined as  $(\phi/\beta)/(1 + (\phi/\beta))$ . For positive  $\phi$ , it is clear that increases in  $\beta$  result in a decrease in  $\gamma$ . In effect, higher values of  $\beta$  are similar, in their effect on the exchange rate, to decreases in the interest elasticity of the demand for money.

In order to complete the model, an assumption must be made concerning the process generating the forcing series,  $z_t$ . For illustrative purposes, assume that the rate of growth of the forcing series is determined by a first-order autoregressive process:

$$(4) \quad \Delta z_t = \alpha \Delta z_{t-1} + v_t.$$

This type of process typically "fits" series like the money supply that are candidates for fundamental variables influencing the exchange rate. Since the procedures for solving equations like (3) with forcing series generated by (4) are well known, only the final results will be reported here. For the exchange rate, the solution is presented in equation (5):

$$(5) \quad s_t = z_t + \frac{\gamma\alpha}{1 - \gamma\alpha} \Delta z_{t-1}.$$

Combining (5) with (2), the forward pre-

mium can be expressed as

$$(6) \quad f_t - s_t = \frac{\gamma\alpha}{\phi(1 - \gamma\alpha)} \Delta z_{t-1}.$$

By combining (6) with (4), the forcing series may be expressed as a first-order autoregressive process:

$$(7) \quad f_t - s_t = \alpha(f_{t-1} - s_{t-1}) + (\gamma\alpha/\phi(1 - \gamma\alpha))v_t.$$

This equation has two important implications. First, it demonstrates that the autocorrelation in the forcing series may be estimated from the autocorrelation in the forward premium. Based upon monthly observations, a reasonable estimate of  $\alpha$  from the floating rate period is about .8 if a constant is included in the equation. Since all of these equations should be considered to be referring to deviations around stable trends, this estimate is appropriate for our purposes. Second, the influence of the failure of the forward parity term on the volatility of the premium can be examined through the term  $(\gamma\alpha)/\phi(1 - \gamma\alpha)$  which translates the innovation in  $z_t$  into the innovation in the premium.

Combining (1), (4), and (5) yields an expression which decomposes the change in the spot rate into its anticipated and unanticipated components:

$$(8) \quad \Delta s_t = \beta(f_t - s_t) + \frac{1}{1 - \gamma\alpha} v_t.$$

With these equations in hand, it is now possible to examine the effect of the failure of the forward parity condition on the volatility of the exchange rate and the forward premium. Using the estimated values of the key parameters, the following estimates of the volatility terms can be obtained.

	$\beta = 1$	$\beta = .2$	Ratio
Exchange Rate: $\frac{1}{1 - \gamma\alpha}$	4.00	4.75	1.19
Forward Premium: $\frac{\gamma\alpha}{\phi(1 - \gamma\alpha)}$	0.2	0.25	1.25

These results can be interpreted in the fol-

lowing way. Under forward parity,  $\beta = 1$ , the standard deviation of the exchange rate will be 4 times the standard deviation of the forcing series. This "magnification effect" reflects the discounting of the change in the anticipated future growth of the series into the current spot price. With  $\beta = .2$ , the market requires a larger premium to accommodate any given change in expectations, the premium itself influences the spot rate through its direct influence on interest rates and the change in the exchange rate is exacerbated. With  $\beta = .2$ , the standard deviation of the exchange rate is 4.75 times the standard deviation of the forcing series. Relative to the forward parity model, the standard deviation of the exchange rate is almost 20 percent higher.

Similar results are obtained for the forward premium. With forward parity, the standard deviation of the premium is estimated to be 20 percent of the standard deviation of the forcing series. With  $\beta = .2$ , this estimate increases to 25 percent. Hence the estimated effect of the risk premium is a 25 percent increase in the variability of the interest rate. In most models of this type, the adjustment to an innovation is shared between the spot price and the premium so that an increase in the variability of one variable should be associated with a decrease in the variability of the other. This is not the case in the present instance.

In the preceding discussion, I have attempted to build a formal link between the empirical results on currency risk premiums and the variability of exchange rates and interest rates. The argument is closely related to Ronald McKinnon's (1979) conjecture that the instability of the floating rate system is due to "insufficient speculation" and to Stanley Black's (1984) exposition of the Harrod effect. Rather than viewing speculation as a source of instability, this literature accepts that speculative activity is stabilizing and argues that the floating exchange rate system inhibits stabilizing speculation. For the remainder of the paper, I will expand upon this theme with some comments and empirical observations.

One useful way to consider the risk premium argument is to view foreign exchange

speculation as an investment activity. William Sharpe (1983) has suggested that the performance of investment advisors can be summarized in the statistic  $\bar{r}/s$ , where  $\bar{r}$  is the average excess return and  $s$  is the standard deviation of the return. For the broad U.S. equity indices, the Sharpe ratio for monthly observations over the postwar period has been estimated to be around .15. Under the assumption that the actual exchange rate follows a random walk, and that the monthly standard deviation is 3 percent, the interest rate differential would have to be 5.4 percent in order to offer the same prior risk/return tradeoff as the market average.<sup>3</sup>

This appears to me to be a wide range. If U.S. interest rates were 10 percent, foreign interest rates could range between 4.6 and 15.4 percent before offering a risk/return tradeoff that was comparable to traditional equity investments. It is also interesting to note that this range encompasses most of the actual interest rates observed on different currencies when U.S. rates were around 10 percent. If major investment institutions required a risk/return tradeoff on foreign currency investments that was at least equal to that on traditional investments, then we would expect that there is a wide range in which interest rates are predominantly determined by domestic factors.

This does appear to be the case. In the absence of currency risk, interest rate differentials should reflect the expected change in the exchange rate. At the other extreme, the interest rate differential between two closed economies should reflect the difference in their expected inflation rates (given similar real rates). A number of studies have demonstrated that nominal interest rate differentials are significant predictors of inflation rate differentials. Discussions in the financial press also invariably attribute changes in U.S. interest rates to domestic, rather than international, considerations.

<sup>3</sup>It is necessary to qualify this risk/return tradeoff argument in two ways. First, the tradeoff may be substantially improved by various portfolio selection techniques. Second, the market or consumption risk of currency speculation may be low relative to the absolute risk.

This argument should not be taken too far. The high nominal interest rates in the United States over the past two years have attracted capital and led to an appreciation of the dollar. The real overvaluation of the dollar has permitted the United States to endure interest rate differentials beyond the 5 percent mentioned above by creating expectations of a future depreciation of the dollar. An interest rate of 8 percent, for example, would attract capital according to the 5 percent rule if the expected depreciation of the dollar was less than 3 percent. This example illustrates why the forward parity regressions yield coefficients that are less than unity. In this case, a 3 percent expected depreciation is associated with an interest rate differential of 8 percent, yielding a  $\beta$  estimate of .38.

The currency risk premium results in a partial segmentation of financial markets. This view may seem to contradict the idea that foreign financial markets offer the opportunity to diversify away purely domestic risk. However, recent estimates by Michael Adler and Bernard Dumas (1983) have shown that the minimum variance portfolio for any individual consists almost entirely of short-term financial instruments denominated in the home currency. In order to be induced to hold riskier investments, the individual must be reasonably confident that the expected return compensates for the risk. There is a great deal of evidence that supports the higher return on risky domestic investments, but there is little evidence to support the view that foreign investments, and particularly foreign currency speculation, has an expected return sufficient to compensate for the risk of the activity.

The currency risk premium is an important factor in considering the relative merits of fixed and floating rate systems. With a fixed exchange rate, financial shocks in a particular country are spread throughout the system and interest rates are linked through the interest rate parity condition. The floating rate system, in contrast, has created a financial environment in which interest rates are predominantly determined by domestic conditions and movements in the exchange rate are dominated by nonspeculative activity. The absence of speculative

activity, particularly activity based upon longer-run fundamentals, may be an important factor contributing to the instability of the system.

I take it to be self-evident that exchange rate and interest rate instability have adverse effects on the world economy. Statistical studies have demonstrated that innovations in nominal interest rates are the primary causal factor (in the Granger sense) in the U.S. business cycle. To the extent that real interest rates have increased to compensate for the risk, the instability has created a painful transfer of wealth between debtors and creditors both domestically and internationally. The combination of real exchange rate instability and real interest rate instability has had particularly adverse effects on the Third World.

A return to a system like Bretton Woods would be one answer to the issues raised in this paper. There are, however, other alternatives that should also be explored. These include tying the value of the dollar to a specific bundle of commodities or creating a new money whose value is determined by an underlying portfolio of real assets. All of these approaches should be considered. For the moment, all that can be done is to recognize that there are many characteristics of the floating rate system that we have yet to understand, let alone control.

## REFERENCES

- Adler, Michael and Dumas, Bernard, "International Portfolio Choice and Corporation Finance: A Synthesis," *Journal of Finance*, September 1983, 58, 925, 984.
- Bilson, John F. O., "Recent Developments in Monetary Models of Exchange Rate Determination," *IMF Staff Papers*, June 1979, 26, 201-23.
- \_\_\_\_\_ and Hsieh, David, "The Profitability of Currency Speculation," manuscript, University of Chicago, 1984.
- Black, Stanley, "The Effect of Alternative Intervention Policies on the Variability of the Exchange Rate: The 'Harrod' Effect," in J. S. Bhandari, ed., *Exchange Rate Management under Uncertainty*, Cambridge: MIT Press, 1984.

Dornbusch, Rudiger, "Expectations and Exchange Rate Dynamics," *Journal of Political Economy*, December 1976, 84, 1161-76.

Fama, Eugene, "Spot and Forward Exchange Rates," manuscript, University of Chicago, 1984.

McKinnon, Ronald I., *Money in International Exchange: The Convertible Currency System*, New York: Oxford University Press, 1979.

Sharpe, William F., *Investments*, New York: Prentice Hall, 1983.

# Reflections on the Exchange Rate System

By J. CARTER MURPHY\*

It is tempting to compare the effectiveness of fixed and flexible exchange rate systems by contrasting the performance of the world economy under the original Bretton Woods rules with economic performance since 1973. Coincidence of events in time, however, does not lead to an inference that the events are causally related. In fact, strong economic performance over a part of the IMF's first twenty-five years and poorer performance since that time are both the results of fundamental forces having little to do with fixed or flexible exchange rates. Here I shall briefly discuss those forces and then comment on proposals that would return us to some form of preannounced official targeting of the rates. This provides me an opportunity to raise some useful questions concerning the functions we want the exchange rates to perform.

## I. The Postwar Boom and Bretton Woods

In a book of a few years ago (1979) I detailed some causes of the breakdown of the original Bretton Woods arrangements, and in the limited space available to me here I must summarize that history. The policy choices of governments were dominated by inflationary biases from World War II to the late 1970's. During the 1950's, inflation was more endemic to Europe and the developing world than to North America, and the effects of the expansionary excesses abroad were offset by periodic devaluations of the inflated currencies against the dollar. When, in the mid-1960's, the United States became, among industrialized countries, the region with greater relative inflation, dollar devaluation against other currencies proved difficult for many reasons, and, in the end, the

gold exchange system centered on the dollar failed when flight from the dollar made the fixed gold-dollar exchange rate untenable. Subsequent efforts to restore currency par values at realigned exchange rates then proved impossible because governmental pledges to support pegged rates lacked credibility. The uneven inflation process was continuing, and inflationary expectations were on the rise.

The crux of the problem was that governments, quite predictably, responded to the constituencies that elected them (and with the limitations of their national political institutions), rather than to the requirements of a particular exchange rate structure whenever internal and external goals conflicted. The ragged inflation that resulted induced massive capital transfers in anticipation of changes in the par value exchange rates. In the end the violence of these transfers wrecked the system's tie to gold and made even continued fixed exchange rates on the dollar untenable for most countries.

To look back now with undue nostalgia on the Bretton Woods arrangements is, I think, a mistake. I like Peter Gray's (1974, p. 16) characterization of the Bretton Woods period as one in which the dollar moved from being an undervalued currency in the late 1940's and early 1950's to one roughly in equilibrium from the mid-1950's to the mid-1960's and finally to a position of overvaluation. In the first of these periods there was widespread discrimination against dollar goods and investments; in the second, progress was made toward trade and investment liberalization; the third saw a strong revival of protectionism in the United States and capital controls abroad.

The disorder that has plagued economic life during the *post*-1973 period of managed floating exchange rates is largely a legacy of the same inflationary excesses that brought the original Bretton Woods rules to an end. Considering older textbook predictions to

\*Department of Economics, Southern Methodist University, Dallas, TX 75275.

the contrary, it is ironic that the great inflation of the 1970's originated in the period of pegged exchange rates while the 1980's policies of disinflation have come in the period of floating rates. The structural maladjustments termed "stagflation" in the early 1970's were exacerbated by ill-fated efforts in Europe and North America to dictate wages and prices and then reinforced by the petroleum crises that began in 1973-74; all these were related to the chronic inflationary management of demand. While the uneven inflation (that supported real economic growth for a time but later became a source of distortion) cannot be blamed on the system of fixed exchange rates, it also cannot be attributed to floating. It was in fact due to economic ideas that concentrated too much on aggregate demand for current output and provided politicians an easy rationale for spending without taxing and for pegging nominal interest rates below equilibrium levels.

## II. Disillusionment with Flexible Exchange Rates

The disorders of the international economy since 1973 have made many economists disillusioned with managed floating and have led some to call for a return to pegging, or at least "targeting," the exchange rates (for example, John Williamson, 1983; Otmar Emminger, 1982; and Atlantic Council, 1983.) Excessive volatility in the rates is said to result from speculative activity based on changing expectations (or in some cases the absence of speculative activity), from lags in price adjustments following exchange rate changes, and from small short-run supply and demand responses to price changes. Protracted misalignments in the rates, an even more serious problem, are attributed to various causes including the following: 1) international capital transfers induced by macro- and microeconomic policies of governments; 2) public and private currency substitutions, motivated by changing appraisals of risks and returns in different currency denominations; 3) lags in sectoral price adjustments; 4) flow-stock mechanics in portfolio adjustments, in which initial portfolio balancing shifts are large relative to more permanent

redistributions of savings flows; 5) asymmetries in market access to information; 6) lagged price and quantity adjustments due to long-term contracts; and 7) oligopolies. (There are reviews of this literature in Williamson, and in Robert Dunn, 1983.) Such a list of complaints begs the question: what do we want the exchange rate system to do?

I believe we want the system: (a) to facilitate an international allocation of resources and pattern of trade that is "efficient"; (b) to encourage the movement of real and financial capital from points of lower to higher value; (c) to deal with systemic risks at low social cost and to assign the burdens of risk taking equitably; and (d) to adjust to disturbances in a low-cost way and to distribute the burden of adjustment costs equitably.

The desire for efficient resource allocation and trade does not imply that the exchange rates should be at purchasing power parity (*PPP*). The *PPP* calculations are behind most allegations of misalignment in the rates even when it is acknowledged that the rates must do more than achieve a goods and services market equilibrium. Williamson goes further than most when he defines a "fundamental equilibrium exchange rate" to be one which is "expected to generate a current account surplus or deficit equal to the underlying capital flow over the cycle, given that the country is pursuing 'internal balance' as best it can and not restricting trade for balance of payments reasons" (p. 14).

Still, what else might the exchange rate accomplish? Do we not want this price to reflect changing expectations as we want other prices to do? It is well established that speculative positions *correctly* taken in advance of uncertain events lead to actions which prepare for the event and have social value. While numerous studies have shown that forward (and, with interest parity, spot) exchange rates are imperfect predictors of future rates and are, to that degree, inefficient, have we experience that an officially targeted rate does better? Even though the "noise" of daily aberrations in a flexible rate is a nuisance, I am skeptical of the view that assumes experts have greater wisdom to gauge market tendencies than market participants have. While I have no difficulty

rationalizing small government interventions to interrupt exchange rate runs or to give depth to thin and hesitant markets, such interventions are very different from exchange rate targeting.

It is clear that exchange rates will "overshoot" their long-run equilibria when exchange rates are free to adjust rapidly while other adjustments in the system are slow. The relevant question is: is such overshooting uneconomic? Overshooting may accelerate other desired responses throughout the system, in the same way that Alfred Marshall (1890) saw his exaggerated short-term price adjustments inducing proper long-term adjustments of supply. Is not every market price in every real world adjustment part of what is at most a "second-best" solution, with the "first-best" solution denied by constraints on rates of adjustment of some variables? Have we much yet to say about the welfare properties of alternative adjustment paths in disequilibrium periods? In view of our ignorance on these matters, I see the evidence on overshooting as useful data on the dynamic properties of national economic systems, but not grounds for rejecting flexible exchange rates in favor of rates more fixed.

Let us turn next to the roles the exchange rate plays in directing real flows of investment. For investment flows to be efficient, it is not required that they be from regions where fixed capital is relatively abundant toward regions where fixed capital is scarce, or from "rich" countries to "poor." The allocation of saving, and also the portion of the existing capital stock that is liquid, is directed by its price and the expected variance of that price. These values are sensitive in different regions to changing time preferences, changing liquidity supply and demand, and changing perceptions of risk, as well as to changes in the efficiency with which investment contributes to the production of goods.

When there are changes in one country's excess demand for liquidity, time preference, or apparent ability to provide safety to capital returns, as compared to other countries, it is, in my view, efficient from a global perspective for capital to move toward the area of

improved or safer returns. One may want to distinguish analytically between changes that originate in the private sector and those that are implemented through government policies—changes in money demand as opposed to government controlled money supply, for example, or changes in aggregate time preference resulting from individuals' savings choices as opposed those due to government managed income redistributions—and I return to this matter below. But I think one should not deny that capital movements induced by changing time preferences, liquidity adequacy, and perceptions of security are "appropriate" and that the exchange rates that accommodate the transfers are efficient rates.

The allegation that recent movements of capital to the United States have sought, among other things, a "safe haven," suggests, however, a further point. The safety sought by these investors has been perhaps in some degree safety from expropriatory governmental regulation. While confiscatory risks are real enough to individuals or firms as investors, they refer to wealth transfers, not losses, from the viewpoint of nations and the world as a whole. If international transfers based on this fear, then, are to be considered socially efficient, the rationale must be in terms of individual liberty and the importance of property rights rather than in terms of other needs.

The subject of risk in the economy, and who should bear it and how, is an especially difficult one, but it is an issue in arguments over the exchange rate regime. When the foreign exchange rate is held within a pre-designated range by official intervention, risks in the system are transferred from some in society to others. Awareness that intervention will prevent a currency's depreciation, for example, relieves those who would be directly affected by such a change (users and producers of internationally traded goods and holders of money and nominal money claims). It also increases risks to others—at home and abroad—including those for whom the likelihood rises that governments will use alternative policy measures to deal with the balance of payments. While the provision of reserve assets for the world economy as a



whole need not be costly (if fiduciary rather than "real" assets are used for this purpose), and even the holding of reserves by one country is costly only in the degree that the reserves are held in assets which yield less return than optimal real investments, use of reserves by any government is costly to those who must forego absorption when the reserve stocks are rebuilt and to those whose absorption is impaired by trade at an officially selected but inappropriate exchange rate. Exchange market intervention by governments, therefore, does not reduce risk; it reassigns it (Milton Friedman, 1953).

A flexible exchange rate spreads adjustment burdens—and the risk of uncertain burdens—*between* countries more than does a stabilized rate because with stabilized rates deficit countries more commonly have to initiate policy changes. On the whole, greater international participation in such burden sharing is probably desirable on grounds of spreading the costs and their attendant risks as widely as possible. One cannot be unequivocal about even this, however, since it can plausibly be argued that adjustment burdens resulting from "bad" policies should be borne in the country responsible for the bad policies, where the policies can be changed.

Business people have deplored since 1973 the increased risks of international trade and investment they have been asked to bear. We are told of the multifaceted adjustments they have made to minimize the costs of their new risks (for example, Marina Whitman, 1984). I do not, however believe that the fact that these burdens and adjustments are being borne by traders and investors is a proper argument against flexible exchange rates. The exchange rate has been described as a "system variable par excellence" (Val Koromzay et al., 1984); it reflects changes that have occurred, are occurring, or are expected to occur in many parts of the world economy. Might it not be best for the uncertainty surrounding these events to be borne, in the first instance, by those most likely to have information about them? In particular, is it not desirable that they be borne to the maximum extent possible by those persons who have less risk aversion and are willing to take

speculative positions in the market, rather than by those who are passive gainers or losers from governments' successes or failures as exchange speculators? Broad price level changes—and their risks and burdens—that accompany significant exchange rate swings are of course passed on by traders in a flexible rate regime; the evidence nevertheless is clear that the effects of many small exchange rate moves are absorbed by the professionals in such a setting. Generally, I feel that risk taking ought to be a market activity, rather than a socialized one, except where significant externalities can be demonstrated.

### III. Which System Best Bears Shocks?

How does the system bear shocks, and can anything be said about the contributions of alternative exchange rate regimes in this connection? Clearly neither fixed nor flexible exchange rates is best at minimizing adjustment costs to *all* types of disturbance. One must, therefore, weigh the seriousness of disturbances the system must absorb and consider what exchange rate design copes best with the more serious shocks. In this vein I suggest that by far the most damaging disturbances to international economic equilibrium during the past forty years have emanated from the public sector, not the private. While many of us were raised in an era in which it was accepted that the *private* sector was the source of economic instability and the *public* sector was the home of policy instruments for countervailing the private sector disturbances, I am sadly driven to doubt the old views. Perhaps a time will come when democratic governments will tax and spend and manage money in a way that is stabilizing to society. But experience since World War II suggests that we should give thought to system designs that absorb shocks from this area as well as the private sector. Furthermore, it is worth noting that to seek a system that protects one national economy from unexpected policy shifts in another is not the same thing as desiring a system that facilitates as much as possible governmental use of the macroeconomic policy instruments.

The ease with which one or another exchange rate regime handles the macroeconomic aspects of public policy disturbances depends, in the first instance, on the way the regime handles the international capital transfers involved, as the asset approach to exchange rate determination has shown us. Now the important thing about a flexible exchange rate regime is that in it net transfers of capital claims immediately become real transfers because international net sales of capital claims affect exchange rates and generate current account surpluses or deficits to mirror the claims flow. In a system where the exchange rate is wholly or partly stabilized, on the other hand, international shifts in capital claims are wholly or partly financed by reserve transfers, avoiding to that degree a disturbance to production and trade. This is, I think, a central argument for stabilized exchange rates, and it is applicable when disturbances are temporary and of sufficiently small magnitude that they can be financed by modest monetary reserve movements. Where a disturbance is large and enduring, a flexible exchange rate is almost certainly preferable because it avoids the production and consumption distortions that are due to a too long maintained nominal exchange rate, and it provides a useful degree of flexibility to real prices and wages at home and abroad.

Since, of course, no one knows whether a government policy change is going to be large or durable until after the event, one must consider the costs involved in a wrong forecast. What is clear here is that the very costly adjustments are the large ones, and they are the ones that must not be missed. Failure to recognize and accommodate them early enough was a critical weakness in the Bretton Woods arrangements. It is this line of thought that persuades me we must have a system among the industrialized countries that has a significant degree of exchange rate flexibility.

Correct expectations and speculation on those disturbances that prove reversible can give even a flexible exchange rate regime the advantages of a fixed rate system. Incorrect expectations or inadequate speculation, on

the other hand, expose the weaknesses in the flexible system. Exploration of crawling rate systems (see my 1965 article) and other compromise arrangements should not be ruled out. But a high degree of exchange rate responsiveness to market pressures is, I believe, an important safeguard against delaying real economic adjustments too long.

#### IV. Containing Public Sector Shocks

It is clear from much of the literature on exchange rate arrangements that the authors seek global adjustments to disturbances that arise in the private sector but minimal external accommodation of disturbances from the public sector. It is not clear to me that all shifts in public sector actions have any less claim to accommodation through world market adjustments than have private sector shifts. Yet the view that one government should not impose the results of its economic policy choices on other nations is widely accepted. It is, I think, the basis for the lament seen in recent economic writings (Whitman; Dunn) that floating exchange rates have not secured isolation for national economies from external changes, and it is the justification for the grant by governments of surveillance authority to the IMF to inhibit beggar-thy-neighbor policy choices. The truth probably is that we don't want bad policies exported, but we do want good policies shared.

Is there any institutional arrangement that, without too great cost, will contain the macroeconomic effects of all government monetary and fiscal changes? Clearly a regime of floating exchange rates will not do so alone, although in a more naive age there was hope for it. Severe restrictions on international capital movements would protect the current account in a flexible exchange rate setting but would be unable to distinguish between government induced capital movements and those serving strictly private sector needs; they would also raise a harvest of other problems. The IMF surveillance of nations' economic policy mixes with a view to discouraging clear beggar-thy-neighbor choices can be helpful, but the IMF's ability to in-

fluence the policies of major countries so far remains small (William Hood, 1982).

One promising guideline for evaluating governments' macroeconomic policy choices calls for fiscal and monetary changes to be "balanced." Equal proportional changes in a nation's monetary base and in face value of public debt instruments outstanding (by maturity class) would go far toward neutralizing the impact of demand management changes on real interest rates in the nation where the changes take place and hence on induced international capital transfers. In a regime of flexible exchange rates, any nation, or all nations simultaneously pursuing such a rule, could follow expansive or restraining overall monetary and fiscal policies of their choice without disturbing the international current and capital account balances much (see my book, 1979). Yet the markets would remain free to accommodate shifts in costs and private preferences of all kinds. Under fixed exchange rates, the monetary-fiscal policy mix of each country is pulled automatically toward the balance of other countries because money stocks move internationally toward countries with the lowest ratio of money to debt growth. But this alone does not prevent policy shifts in one country from impinging upon others.

I do not hold out much hope in the near future for guidelines calling for fiscal and monetary balance because fiscal policies are the result of microeconomic political pressures as much as of macroeconomic needs, and central banks will, for reasons, some of which are good, be reluctant to yield up their independence to pursue their own policy objectives. Nevertheless, if we continue to have activist government policies and at the same time wish to limit their impact abroad, we must create new images for the public concerning what constitutes good policy. In this case, flexible exchange rates with monetary and fiscal policy balance provides one of the few packages around that is deserving of consideration.

## REFERENCES

- Dunn, Robert M., Jr., *The Many Disappointments of Flexible Exchange Rates*, Essays in International Finance, No. 154, Princeton University, 1983.
- Emminger, Otmar, *Exchange Rate Policy Reconsidered*, Occasional Papers 10, Group of Thirty, New York, 1982.
- Friedman, Milton, "The Case for Flexible Exchange Rates," in his *Essays in Positive Economics*, Chicago: University of Chicago Press, 1953.
- Gray, H. Peter, *An Aggregate Theory of International Payments Adjustment*, Lexington: D. C. Heath, 1974.
- Hood, William, C., "Surveillance Over Exchange Rates," *Finance and Development*, March 1982, 19, 9-12.
- Koromzay, Val, Llewellyn, John and Potter, Stephen, "Exchange Rate Rules and Policy Choices: Some Lessons from Interdependence in a Multilateral Perspective," *American Economic Review Proceedings*, May 1984, 74, 311-15.
- Marshall, Alfred, *Principles of Economics* (1890), 8th ed., London: Macmillan, 1946.
- Murphy, J. Carter, *The International Monetary System: Beyond the First Stage of Reform*, Washington: American Enterprise Institute, 1979.
- , "Moderated Exchange Rate Variability," *National Banking Review*, December 1965, 2, 151-61.
- Whitman, Marina v. N., "Assessing Greater Variability of Exchange Rates: A Private Sector Perspective," *American Economic Review Proceedings*, May 1984, 74, 298-304.
- Williamson, John, *The Exchange Rate System*, Washington: Institute for International Economics, 1983.
- Atlantic Council, "Policy Paper on the International Monetary System: Exchange Rates and International Indebtedness," Working Group on International Monetary Affairs, Washington, 1983.

# On the System in Bretton Woods

By JOHN WILLIAMSON\*

It has become customary to look back with nostalgia at the golden age when the Bretton Woods system held sway, from the early 1950's to around 1970. It also seems to be the conventional wisdom, however, that the rules of Bretton Woods contributed little to the impressive performance of the world economy over that period—a performance characterized not merely by the fastest and most widely distributed growth in history, but also by notable stability, including near price stability except at the beginning and end of the period. The deterioration in the performance of the world economy since the early 1970's is viewed as a coincidence, or a response to common causes, rather than as a consequence of the breakdown of Bretton Woods. My purpose in selecting the title to this session was to induce critical scrutiny of these conventional attitudes.

My own contribution to this task will start by describing what I conceive to have been the three essential rules of the Bretton Woods system. I shall proceed to examine the logic of those three rules in terms of recent contributions to the literature, in particular the emergent literature on policy coordination, and of recent historical experience.

## I. Three Central Rules of Bretton Woods

An international monetary system comprises an exchange rate regime, rules governing the policies that are to be used to adjust payments imbalances, and a reserve supply mechanism to provide assets in which payments imbalances are settled.

Bretton Woods adopted the adjustable peg. *Exchange rates* were normally to remain sta-

ble within narrow margins, but could, and by implication should, be changed when there was a "fundamental disequilibrium." Although never formally defined, and therefore a target for much critical academic comment, there was never much doubt what this concept meant: a situation in which a country could not expect to achieve basic balance over the cycle as a whole without deflating output from full capacity or restricting trade or payments for balance of payments reasons. Thus the basic principle embraced was that exchange rates should be directed toward medium-run balance of payments needs rather than short-run anticyclical policy. The motivation was quite explicitly that of outlawing the beggar-thy-neighbor use of exchange rate policy that had occurred in the 1930's.

The IMF Articles adopted at Bretton Woods were not comprehensive in their specification of the practices that were to govern *balance of payments adjustment*. For example, there was no explicit specification of which country should initiate adjustment, or when. Exchange rates could be adjusted when there was a significant medium-run imbalance. Reserves were to be used to avoid the need for continuous external balance. It was universally assumed at the time of Bretton Woods, though not formally spelled out, that fiscal and monetary policy would be directed at the maintenance of full employment (or "internal balance"). But by the absence of alternative provisions to deal with modest non-self-reversing imbalances, one infers that the architects of Bretton Woods accepted that it would be necessary to shade fiscal-monetary policy with a view to the balance of payments position.

Although Keynes did not get the negative interest rates on creditor *bancor* positions that he sought, the *reserve regime* did imply something about the assignment of adjustment responsibilities. Countries were expected to restrict their deficits to the sums

\*Senior Fellow, Institute for International Economics, 11 Dupont Circle, NW, Washington, D.C. 20036. I acknowledge constructive comments on a previous draft by Matthew B. Canzoneri, William R. Cline, Michael Jones, Stephen Marris, and Robert Solomon; the usual caveat applies. © Copyright 1985 Institute for International Economics. Published here by permission.

that could be financed from their available reserves supplemented by IMF drawing rights. Reserve currency countries (*sic*) had some additional latitude to finance payments deficits by issuing their own currencies, but it was assumed that this possibility would be limited by the need to maintain confidence in convertibility. Bretton Woods ratified the gold-exchange standard, it did not legislate a dollar standard. (The special position of the dollar was confined to the obligation to defend a par value in terms of gold rather than in terms of another currency.) Britain invested much energy in gaining an assurance that the United States would participate in the adjustment process (and secured the scarce currency clause to that end), which would have been anomalous in a dollar standard world.

In my interpretation, Bretton Woods was far more than just a commitment to pegged exchange rates, as has sometimes been claimed (for example, see Ralph Bryant, 1980, p. 475; Robert Solomon, 1984, p. 174). It embodied, rather, a comprehensive set of rules for assigning macroeconomic policies: exchange rates to medium-run external balance, fiscal-monetary policy to short-run internal balance, and reserves to provide a buffer stock (as distinct from a monetary base) that would allow short-run departures from external balance. This is the intellectual position that Keynes had developed in the interwar years, which in my view probably explains why he used his influence to secure British ratification of Bretton Woods despite his bitterness at the across-the-board rejection of the plans for postwar reconstruction that he had nurtured during the war years (see my 1983a article).

## II. Implications of the Bretton Woods Rules

A vast literature developed in the 1960's on the ills of the Bretton Woods system. These were classified under the headings of the problems of adjustment, liquidity, and confidence. This is not the place to review those topics, save to say that this classification became such a part of the conventional wisdom that it has sometimes been used as a framework within which to evaluate *any* in-

ternational monetary system. This is misguided. One may agree that adjustment, liquidity, and confidence are all handled better by present arrangements than they were by the Bretton Woods system, but nevertheless believe that *other* problems with present arrangements outweigh those gains. What seem to me the dominant problems with these arrangements are (a) their propensity to generate exchange rate misalignments, (b) the absence of any discipline on national overspending less drastic than the credit-worthiness constraint that has now replaced the liquidity constraint, and (c) the lack of pressure they exert to coordinate policies. One may ask, with the benefit of hindsight, why these were *not* problems with the Bretton Woods system.

*Exchange rate misalignments* remained modest for most of the Bretton Woods period. The first rule of the system said that a misaligned rate (using that term as a pseudonym for fundamental disequilibrium) could be adjusted to eliminate the misalignment. One objective of the advocates of limited flexibility was to transform this *right* to change a misaligned rate into a positive *duty*. In the absence of such a duty, and in the presence of the pressures that the adjustable peg created to declare rates permanently fixed (so as to discourage speculative attacks), some substantial misalignments—especially of the reserve currencies—emerged by the second half of the 1960's. However, even these misalignments were smaller than those witnessed in recent years. Misalignments emerged from differential inflation and productivity growth, but not from exchange rate movements: indeed, since markets knew that governments would change par values only when they judged this appropriate on the basis of long-run fundamentals, speculative pressures acted as a check on the size of misalignments, rather than as a cause of them.

There is in my view a presumption that the vast exchange rate misalignments of recent years have had some negative effect on growth (via the promotion of deindustrialization in countries with overvalued currencies) and also some effect in ratcheting up inflation (see my 1983b study). These effects are,

admittedly, not well documented, though the supposed negative tests of the ratchet hypothesis are unpersuasive since they did not search for a ratchet effect on wages.

The virtues of a *liquidity constraint*, which resulted from the limited mobility of capital, were not always appreciated during the days of Bretton Woods. As recent work by Michael Jones (1983) has shown, however, a reserve constraint can substitute for explicit policy coordination, a theme discussed further below. The social function of a reserve constraint is there interpreted as that of disciplining macroeconomic policies so as to limit and make more predictable the demand and monetary spillovers received from abroad. In a world where policymaking is often myopic, the displacement of the former liquidity constraint by a creditworthiness constraint is not necessarily advantageous even from the standpoint of the individual country involved. Under Bretton Woods, countries suffered speculative attacks when their par values came to look vulnerable, instead of being able to pursue unsustainable policies for years while building up a vast mountain of foreign debt. That ability has already led Latin America into a devastating debt crisis, and it is now permitting the United States to pursue a policy course equally fraught with danger. Extensive capital mobility on the scale of the past decade is like fiat money: a social innovation with potential for great good, which in fact has almost certainly detracted from human welfare.

Overt (sometimes described as "continuous") *policy coordination* does not seem to have been much more effective in the golden age of Bretton Woods than it is in the days of Reaganomics. Hence the claim that Bretton Woods secured a useful measure of policy coordination has to rest on a demonstration that it was a successful case of what Richard Cooper calls a "rule-bound regime" (1985, p. 1226). That is, that the Bretton Woods rules were such as to ensure that spontaneous pursuit of national self-interest subject to continued observance of the rules ensured a broad consistency between national and world interests.

There is by now a significant literature showing formally that choice of monetary

and fiscal policies with a view to short-run stabilization objectives (maximization of a utility function specified in terms of the activity level and inflation and/or the balance of payments on current account), taking other countries' policies as given and with exchange rates flexible, leads to an outcome (the Nash equilibrium) inferior to that available with policy coordination (the cooperative solution). See Matthew Canzoneri and Jo Anna Gray (1983), David Currie and Paul Levine (1984), and Gilles Oudiz and Jeffrey Sachs (1984). The first two of these three contributions investigate explicitly the merits of a fixed exchange rate rule as a way of ruling out manipulation of the monetary/fiscal mix with a view to gaining short-run benefits at the expense of other countries, and conclude that this constraint can indeed be mutually beneficial. The first of the Bretton Woods rules introduced precisely this constraint, but in a form that did not preclude an exchange rate change needed to facilitate payments adjustment.

The objective of precluding the manipulation of exchange rates as instruments of short-term anticyclical policy was indeed deliberately sought by the architects of Bretton Woods, with the experience of competitive devaluations in the 1930's in mind. Has experience since 1973 shown comparable evidence of antisocial behavior? To answer that question one has to know whether it is appreciation or, as in the 1930's, depreciation (relative to the *PPP* trend) that will have an adverse impact on other countries. Canzoneri and Gray have postulated that it is structural characteristics in the world economy, such as the degree of wage indexation and whether the oil price is fixed in terms of a particular currency, which determine whether the spillover benefits of monetary expansion on other countries are positive or negative. My own hypothesis would be simpler; that this depends primarily on whether unemployment or inflation is perceived to be the dominant problem confronting policymakers. In the 1930's, depreciation was antisocial. But when inflation is perceived to be the dominant problem, as in general since 1973, complaints will center on competitive appreciation—as they did in the debate on vicious and virtu-

ous circles in the mid-1970's, and in the reaction of Europe and Japan to the dollar appreciation of the early 1980's. The recent muting of European complaints on this score may reflect growing recognition that the problem of the decade in Europe is unemployment.

The second of the Bretton Woods rules as interpreted above assigned monetary-fiscal policy to short-run internal balance, subject to the need to maintain the exchange rate peg. As long as the basic presupposition of Keynesian demand management, that price levels are constant (or at least predetermined), remained approximately valid, this worked rather well: with hindsight, fine-tuning surely has to be rated a greater success than it was at the time. And as long as exchange rates remained properly aligned, the "stop-go" policies imposed on demand management by the Bretton Woods constraints served a social function that few appreciated at the time (but see Robert Triffin, 1960, pp. 82-83, for an exception). With a correctly aligned exchange rate, an external deficit provides an early warning that deflation is needed, while the automatic stabilizer provided by the deflection of excess demand into an external deficit constitutes a safety valve that helped prevent the development of inflationary inertia.

The third of the Bretton Woods rules required countries to respect a (reasonably symmetrical) reserve constraint, while using reserves as a buffer stock to reconcile continual pursuit of internal balance with only medium-run pursuit of external balance. The major benefit that arose from this rule, in conjunction with the first two, is that it helped to avoid a synchronized world business cycle. Robert Lawrence (1978) found that there was practically no synchronization during the period 1959-67 when the Bretton Woods system was functioning properly, with adequate but not excessive liquidity. There was synchronization in an earlier period, 1950-58, which he explained by a liquidity shortage that forced other countries into following the United States, especially during and after the Korean War. The recession of 1958 was the one case of a synchronized (but nonetheless relatively mild) world recession during

the Bretton Woods period. Thereafter until 1969 recessions were essentially national, induced primarily by a need to curb payments deficits: France in 1959; the United States in 1960; Britain and Canada in 1962; Italy in 1963-64; Japan in 1964; Germany in 1966. Under the Bretton Woods rules, other countries tended to offset those recessions by expanding demand so as to preserve internal balance when exports fell, while one country's deficit (that pushed it toward restraint) was another's surplus (which nudged it toward monetary expansion). Thus the world as a whole did not deviate significantly from full employment because of payments imbalances. The mechanism that was supposed to secure this result under the gold standard (though it is doubtful whether it did—Barry Eichengreen, 1984) seems actually to have functioned during the golden age of Bretton Woods. Advocates of a dollar standard, such as Charles Kindleberger (1981), never gave sufficient weight to the loss of this mechanism that would follow abandonment of a U.S. reserve constraint. (That constraint *did* have an impact on U.S. policy: witness John Kennedy's nightmares about a rise in the price of gold, Operation Twist, and the 1968 tax increase.)

The evidence of Lawrence and Alexander Swoboda (1983) reinforces casual retrospection and an examination of the reference cycles of the Center for International Business Cycle Research in confirming that there has, in contrast, been a pronounced world cycle since 1969. The only question is the extent to which this can legitimately be attributed to the breakdown of Bretton Woods. Those of us who date the beginning of the end of Bretton Woods to the abandonment of the gold pool in March 1968, on the grounds that this transformed the reasonable symmetry of Bretton Woods' reserve constraint into the asymmetry of a dollar standard which paved the way for "benign neglect," may argue that the inflationary boom of the early 1970's can be explained that way. Doubtless the oil price increases of 1973 and 1979 were the primary causes of the ensuing global recessions, but the reason that both recessions were more severe than policymakers would have chosen *ex ante* on

grounds of anti-inflation policy can most plausibly be explained as a consequence of failure to appreciate the cumulative impact of other countries' policies (Lucio Izzo and Luigi Spaventa, 1981) and of attempts to engineer "virtuous circles."

### III. Concluding Remarks

It has often been said by supporters of present arrangements that international monetary problems derive from the policies countries pursue, not from "the system." If only each country would look after its own fundamentals, the system would look after itself. The converse of that proposition is undoubtedly true: if countries do not take due account of fundamentals, including fundamentals that produce ill effects only in the long run (if they pursue unemployment targets below the natural rate, or accumulate debt in quantities that jeopardize creditworthiness), the system is in trouble. But the proposition itself seems to be only as valid as the thesis that short-run anticyclical policies are useless. To the extent that there is a role for demand management policy to react to shocks, it is better that countries select their policies under the constraint that the (real) exchange rate not be changed except for medium-run adjustment. By imposing that constraint in the context of a system of reasonably symmetrical reserve constraints and an expectation that countries would aim for full employment, Bretton Woods contributed significantly to the stability, and therefore to the longevity, of the postwar boom. It accomplished this by influencing the policies that countries pursue: the antithesis between countries' policies and the international monetary system is a false one. To say that existing arrangements are the only ones that could have reconciled the policies pursued over the past decade is to condemn those arrangements, not to dismiss alternatives.

To recognize that the architects of Bretton Woods sought a comprehensive and enlightened policy assignment (and I cannot otherwise make sense of the wartime discussions), and that this assignment did indeed guide policy in the postwar world, is not to overlook the deep flaws in the Bretton Woods

system. One can argue that the system worked as well as it did for a little over a decade because of a series of happy accidents: liquidity was about right despite the absence of any mechanism to secure that result, exchange rates were reasonably aligned despite the attempt to suppress realignments, inflation was low, the reserve center still acted as though the suspension of gold convertibility would be a disaster. Bretton Woods did not provide mechanisms capable of maintaining such a satisfactory conjunction of circumstances. The SDR agreement was intended to provide that flexibility with regard to the quantity of reserves, but it was not complemented by those additional reforms (limited exchange rate flexibility, asset settlement) that were needed to allow the system to survive. My contention is that the neglected virtues of Bretton Woods were sufficiently real to make that failure a matter of regret.

Bretton Woods worked for a while because its rules were consistent with the needs of the time and, until emergence of the dollar overvaluation, acceptable to the dominant power. Some of its rules, such as an orientation of exchange rates to medium-run objectives, could be a useful element of any return to an organized system. In other respects it is clear that the rules would have to be very different: there is no prospect of legislating a return to effective reserve constraints. But those who are shocked by the complacency with which Washington is viewing the current dollar overvaluation must agree with the political scientists Robert Keohane and Joseph Nye (1984, p. 25), as well as Cooper (pp. 1227-28), who argued that international regimes are to be sought to protect countries not only from competitive behavior by their peers, but also from their own short-sighted follies.

### REFERENCES

- Bryant, Ralph C., *Money and Monetary Policy in Interdependent Nations*, Washington: The Brookings Institution, 1980.
- Canzoneri, Matthew B. and Gray, Jo Anna, "Monetary Policy Gains and the Consequences of Non-Cooperative Behavior,"



- International Finance Discussion Paper No. 219, Federal Reserve Board, February 1983.
- Cooper, Richard N., "Economic Interdependence and Coordination of Economic Policies," in Ronald Jones and Peter B. Kenen, eds., *Handbook of International Economics*, Vol. II, Amsterdam: Elsevier, 1985.
- Currie, David and Levine, Paul, "Macroeconomic Policy Design in an Interdependent World," paper presented to the Conference on the International Coordination of Economic Policy, Centre for Economic Policy Research, London, 1984.
- Eichengreen, Barry, "International Policy Coordination in Historical Perspective: A View from the Interwar Years," paper presented to the Conference on the International Coordination of Economic Policy, Centre for Economic Policy Research, London, 1984.
- Izzo, Lucio and Spaventa, Luigi, "Macroeconomic Policies in Western European Countries, 1973-77," in H. Giersch, ed., *Macroeconomic Policies for Growth and Stability: A European Perspective*, Tübingen: J. C. B. Mohr, 1981.
- Jones, Michael, "International Liquidity: A Welfare Analysis," *Quarterly Journal of Economics*, February 1983, 98, 1-23.
- Keohane, Robert O. and Nye, Joseph S., "Beyond Dreams of World Government," mimeo., Harvard University, 1984.
- Kindleberger, Charles P., *International Money*, London: Allen and Unwin, 1981.
- Lawrence, Robert Z., "The Measurement and Causes of the Synchronization of the International Business Cycle," unpublished doctoral dissertation, Yale University, 1978.
- Oudiz, Gilles and Sachs, Jeffrey, "Macroeconomic Policy Coordination among the Industrial Economies," *Brookings Papers on Economic Activity*, 1:1984, 1-76.
- Solomon, Robert, "Discussion" of paper by Robert Triffin, in *The International Monetary System: Forty Years After Bretton Woods*, Boston: Federal Reserve Bank, 1984.
- Swoboda, Alexander K., "Exchange Rate Regimes and U.S.-European Policy Interdependence," *IMF Staff Papers*, March 1983, 30, 75-102.
- Triffin, Robert, *Gold and the Dollar Crisis*, New Haven: Yale University Press, 1960.
- Williamson, John, (1983a) "Keynes and the International Economic Order," in D. Worswick and J. Trevithick, eds., *Keynes and the Modern World*, Cambridge: Cambridge University Press, 1983.
- \_\_\_\_\_, (1983b) *The Exchange Rate System*, Washington: Institute for International Economics, 1983.

## ECONOMIC EDUCATION: THE USE OF COMPUTERS<sup>†</sup>

### Computer Applications in Pre-College Economics

By JOHN M. SUMANSKY\*

Despite the fact that computer technology has become commonplace in the schools, its full impact on education in general and economics education in particular has yet to materialize. Some have been disappointed, expecting that the impact would have been felt by now. Others take the fact that the impact has not yet materialized to signify that it will never occur. Still others suggest that the true educational impact is now only beginning to unfold and may as yet not be able to be known by us. Although the temptation is great to explore these positions more fully; rather, in this paper I want to accomplish somewhat more modest objectives.

First, I will provide a sketch of the landscape of present day economics instruction in the nation's schools. Against this landscape, I will then discuss the software which is available for the teaching of economics, K-12. This discussion will lead to at least a partial explanation of the general dissatisfaction with the present generation of software available for economics instruction.

Second, I will discuss what, in my opinion, is a major obstacle limiting the ability of economics to take full advantage of the computer. My objective will be to identify action that must be taken if economics is to keep pace with other disciplines in developing effective educational uses of the computer.

#### I. The Landscape: Reality of Pre-College Economics Instruction

At the pre-college level, the microcomputer has more or less been superimposed on the existing educational landscape rather than having had a new one painted in which the computer was an integral part of its composition. To be able to judge how successful this imposition has been, we need to develop an understanding of the existing landscape by looking at several of its features.

First, in the teaching of economics at the pre-college level, one has to grapple with cognitive constraints. Early age students are less able to deal with complex ideas, abstractions, and higher-order thinking than their older counterparts. This fact constrains what and how much economics *can* be taught.

Another important feature relates to the curriculum. There is a core curriculum, namely reading, writing, and arithmetic. As presently structured, economics is not likely to be found as a separate area of study. If found, it is an integrated part of some other subject matter.

A third feature is directly related to the centrality of the textbook in pre-college instruction. What economics is taught, is overwhelmingly related to the textbook. If the text in use covers an economic topic, then teachers likely will cover it in the classroom.

A fourth feature of the K-12 education landscape has to do with the presence of the computer. Today nearly 85 percent of schools in this nation have computer capabilities and that percentage is expected to rise to 100 percent over the next few years as the price of hardware declines. The predominant hardware in use is the Apple, with TRS-80 and Commodore also in evidence. These three

<sup>†</sup>*Discussants:* Susan K. Feigenbaum, Claremont McKenna College and Claremont Graduate School; Herbert R. Fraser, H. W. Fraser Associates; F. Trenery Dolbear, Brandeis University.

\*Program Director, Joint Council on Economic Education, 2 Park Avenue, New York, NY 10016. Thanks are extended to Robert Highsmith and Michael MacDowell for helpful comments.

account for about 80–90 percent of all hardware presently in use.

Although the landscape contains a number of other dimensions, the four above features define the picture well enough for present purposes. And it is against this landscape that I offer some general observations about the software that is known and available for teaching economics at the pre-college level.

The list of known and available software for teaching economics has grown to more than 150 titles—covering economics instruction for all grade levels and ranging over a myriad of topics from coin-counting to macroeconomic simulations. The vast majority are written for use on the Apple.

The software is not distributed equally across the grade levels. More is available for use at the upper grades than at the lower grades, a fact which is consistent with the grade location of computers and with the curriculum.

At lower-grade levels, software is available to teach single, simple concepts at cognitive levels appropriate to the age level. The predominant computer applications are drill and practice and tutorial, both of which are appropriate to the teaching/learning environment in the lower grades. For example, the Joint Council on Economic Education's *Piggy Bank* (JCEE, 1984), a software package for K-3rd grade students, introduces the abstract concept of money in concrete terms, a presentation which is appropriate for 7–8 year-old students. And the learning activity is presented in a curriculum context familiar to teachers, that is, developing basic math skills such as counting.

At the high school level, software packages are available dealing with complex concepts and topics. For example, the JCEE *Marketplace* is a three-cognitive-level tutorial on price determination. The complex concept is treated abstractly using approaches which are consistent with the intellectual capacity of the high school student, textbook, and curriculum.

Experience—as evidenced by software developed to date—demonstrates that the computer can be programmed to produce learning activities fairly easily. Individuals have

been successful in developing imaginative and interesting economics activities, when judged against the landscape of pre-college economics instruction. While not all software for economics instruction have desirable characteristics, enough do that they should cause us to be optimistic about the future.

In addition to optimism arising from mild successes to date, there is another major reason for optimism about the future of computers and economics education. The structure of economic knowledge, the ability to break that knowledge down into component parts, the cumulative nature of acquisition of economic knowledge, and the ease of finding examples of ways of usefully applying economic knowledge, all have great consistency with the way in which knowledge must be structured and interpreted in order to develop computer learning activities.

Despite my fairly generous and optimistic views on the present generation of economics software, it is difficult to ignore the literature on computers in the schools which contains more than passing reference to the inadequacy of education software in general. Much the same has been heard about economics software at the K–12 level.

In my opinion, there are two main reasons for the general dissatisfaction. One reason stems from the fact that appropriate evaluation paradigms for software/computer use have yet to be developed. Especially at the grade school level, “it is difficult to test for the effectiveness of a simulation package in improving educational attainment when that package attempts to achieve goals which are difficult to achieve with a textbook, lecture or discussion” (Robert Schenk and John Silvia, 1984, p. 241). Research on the effectiveness of the microcomputer as a substitute for or complement to present teaching and learning—especially at the K–12 levels—is nonexistent! In its place is uncertainty and uncertainty breeds dissatisfaction.

Another and perhaps more fundamental dissatisfaction with present software arises from a widespread belief that “drill and practice” and “tutorials” are glorified page turners. Since observers find most K–12 software to be drill and practice and tutorial

aimed at lower cognitive levels, they naturally judge the present generation of software to be poor. I part company with those who base their judgement on this criterion because there is "...considerable evidence that intellectual skills...require considerable exercise before they become fully effective. Consequently, even though practice is a less glamorous goal for computer-based education, it is one that should be vigorously pursued" (Robert Lesgold, 1983, p. 69).

I think the criticism of and dissatisfaction with present-day educational software is serious, *but* not because drill and practice and tutorial are inherently poor uses of the computer, but rather because the dissatisfaction reflects a much more serious problem. In Section II, I will argue that a new body of research knowledge will be required to enable the discipline of economics to take full advantage of the computer.

The absence of research of the kind needed has forced software developers to rely on personal classroom experience and professional intuition to design instructional activities on the computer. Since experiences and intuition differ from one developer to the next, we tend to see rather different paths being mapped out to reach identical learning outcomes. This instructional design technique is certainly not unique to the preparation of computer materials. One need only pick up any two principles texts to see two authors differing in their views about what ideas in what sequence are prerequisite to learning a complex concept. But, unlike printed texts, a user other than developer cannot alter the program to suit his or her own "style" or approach. So, the program either is dismissed as incorrect, inappropriate, or interesting but trivial.

## II. The Computer and Economics: Potential and Path to Progress

Up until now, the more powerful applications of the computer have eluded economics and will continue to do so until research of a particular kind is conducted which will *enable* developers of economics software to produce computer activities which demand intensive intellectual participation on the part of the

student, guide students through *optimal* sequences of discrete pieces of economic knowledge accumulating to a desired learning objective, diagnose and treat errors in reasoning, and build on the unique characteristics of individual students. This is the potential of the computer, a potential which is as yet beyond the grasp of economics.

For progress to be made in designing powerful computer applications in economics, it is clear that intuition, personal observation, and the simple transformation of unique classroom lectures into computer lectures will not serve us well in the future. What is required is the filling of a research agenda which was outlined exactly thirty years ago at the Riverdale Conference. The conference issued the following call:

Economists and educators alike need to know what problems their students—of varying abilities—*are* interested in; or *can be* interested in. They need to know whether the commonly used methods for working through problems to understanding *actually* develop the knowledge and competence that are hoped for. They need to understand the learners' environment against which things to be learned are by necessity presented.

Work of a realistic, perceptive, and intelligent sort depends upon an accurate and complete knowledge of facts (what is being done, what can be done, the relationships among various factors and teaching success however defined), it seems undesirable to trust "common sense", or unsystematically acquired "experience", for its acquisition. Systematic, objective, scientific investigation—a program of research—seems definitely to be called for.

[James Gemmell et al., 1954, p. 142]

While the Riverdale Conference did not deal explicitly with computer applications, the questions raised there, in retrospect, are most relevant to the task of producing knowledge that will enable the development of more powerful economic education software. In fact, had their suggestions been followed, economics software might have had a different look today. As it stands, econom-

ics is not included among those disciplines that have a robust understanding of the role of lower cognitive skill development in leading to higher skills. For example,

In mathematics, research efforts are beginning to achieve a rapprochement between understanding and drill; in reading evidence accumulates for the dependence of the "intellectual" aspects of reading on automation of lower level skills, such as word recognition; even in domains such as radiology, the role of highly overlearned perceptual skills in the course of diagnostic reasoning is becoming important. [Lesgold, p. 69]

Because of this research, other disciplines are much more comfortable with and accepting of drill and practice and tutorial programs than economics seems to be. And their drill and practice and tutorials are quite different than anything available in economics. In short, economics does not have a literature comparable to mathematics and physics. Witness the following sample of paper titles: "Unlearning Aristotelian Physics: A Study of Knowledge Based Learning"; "Expert and Novice Performance in Solving Physics Problems"; and "Learning Without Understanding: The Effect of Tutoring Strategies on Algebra Misconceptions." A quick perusal of this list suggests that research in other disciplines has been and is being pursued to answer three important questions: is there anything "systematic" about errors students make? What are the key differences between "expert" thought/problem solving and "novice" approaches? What are the component parts (i.e., the building blocks) of the disciplines in question? These issues bear a strong resemblance to those identified at the Riverdale Conference.

It is my contention that answers to these questions are basic to the development of educational software of great value. I argue that because mathematics and physics and several other disciplines have such a literature, they also have a "different" type of software available.

In physics, for example, computer programs are available that allow students to

compare what would happen if their naive beliefs were true with what happens in a world governed by Newton's laws. This program would not be possible *unless* their research program clarified differences in novice knowledge vs. expert knowledge.

In mathematics, a programmed diagnostic system is available that determines whether deleting one or two specific steps in the subtraction process leads to the exact error pattern a child displayed. The program allows a specific conceptual problem to be more readily identified and then overcome. It is possible not only to detect missing knowledge, but also to provide hints that lead students to discover missing steps in their mathematics procedures.

This program—though far from perfect—would not have been possible without a prior research program which identified the specific steps in mathematics problem solving and systematically linked errors in math to the missing or incorrect steps.

A number of other examples could be given here. But, in general, it is likely that the absence of economic education research of the kind being done in several other disciplines will hinder the ability of the economics discipline to take full advantage of the power of the computer. Other disciplines are making rapid advances along these lines, economics is not.

### III. The Future: A Broadened Research Agenda

There are some immediate needs for research on the use of the computer in economics. There is an immediate need for answers to questions like:

1) Is the computer more effective as a substitute for or complement to alternative means of transmitting economic knowledge?

2) To what extent does drill and practice activity contribute to higher-order learning in economics?

3) What are the optimal classroom environments and educational inputs for using the computer in teaching economics?

A longer-term research program needs to begin to identify how economic knowledge can be broken down into its component parts and how the parts are to be assembled in

sequences leading to specific learning outcomes. Prerequisite knowledge, including the experience and intellectual maturity of students at various age levels, needs to be addressed against the backdrop of the cognitive processes of information accumulation, information processing, and information utilization. Each of these cognitive processes needs substantial research to uncover its marginal contribution to "expert" economic knowledge levels in economics.

For example, with the process of information accumulation, just what role do personal experiences play, how important to higher-order thinking are personal experiences, and how important to higher-order thinking is knowledge of the definitions of key economics concepts?

Research needs to be done on information processing—that is, what cognitive processes or problem-solving models do students use to sift through economic information?

We need to determine not only which ideas and concepts are prerequisite to higher-order thinking in economics, but also how one goes about recognizing them when they are not known by novices, and how one knows when to reach out and draw in another concept and how to link them together in a sequence to be able to deal with economics problems/issues of high complexity.

Finally, differences between "experts" and "novices" in the ways in which economic information is accumulated, processed, and utilized needs to be studied. Perhaps this is the most crucial item on the research agenda.

This is an imposing research agenda, yet an important one if we are to truly realize the potential of the computer. And, while it is the quest for realizing the educational potential of the computer which heightened the urgency of this research, the computer also appears to be a possible solution to some of the questions raised. The computer appears to have great power as a research tool—because it offers researchers a way of presenting alternative economic learning se-

quences without being tainted by instructor differences.

Experiments using alternative sequences can be devised to measure the differences in learning outcomes for students with different backgrounds, experiences, intellectual maturity, prior knowledge, and learning styles. The ability to replicate such studies is also enhanced.

Thus, intuition and experience no longer need be the only information used to decide which knowledge sequence is best for teaching a particular economic concept. Armed with this new knowledge, "smart" programs can be written to diagnose student readiness for economics learning, select the best sequence of ideas, and present the student with virtually an unlimited array of remedial work and application problems relevant to the student's characteristics.

## REFERENCES

- Gemmell, James, Harris, Seymour and McCutchen, S. P., *Economics in General Education: Proceedings of the Riverdale Conference*, New York: Joint Council on Economic Education, 1954.
- Lesgold, Robert, "Paradigms for Computer-Based Education," in *Computers in Education: Realizing the Potential*, U.S. Department of Education, Washington: USGPO, August 1983.
- Schenk, Robert and Silvia, John, "Why Has CAI Not Been More Successful in Economic Education? A Note," *Journal of Economic Education*, Summer 1984, 15, 239-42.
- Joint Council on Economic Education, *Marketplace (Grades 10-12)*, Version 2.0, (Economic Education Software Sampler Series, Unit 5), New York, JCEE, 1984.
- \_\_\_\_\_, *Piggy Bank (Grades K-3)*, Version 2.0, (Economic Education Software Sampler Series, Unit 1). New York: JCEE, 1984.

# Macro Simulations for PCs in the Classroom

By KARL E. CASE AND RAY C. FAIR\*

There has always been a large difference between macroeconomics in the classroom and macroeconomics as it is used in practice. Macroeconomics of the classroom is basically theoretical, and it is taught almost exclusively in a simple comparative statics framework. Practicing macroeconomists, on the other hand, work with time-series data, sophisticated statistical techniques, and large-scale macroeconometric models. The PC revolution has now made it possible to bring macroeconometric models into the classroom.

The model in Fair (1984), which consists of 128 equations, has been programmed to run on a PC; this paper will discuss its potential for use as a teaching tool at the elementary and advanced levels. The experience so far has been quite encouraging.

## I. Hardware Requirements

The software was written with the aim of minimizing the amount of memory needed to run the program. This amount turned out to be 128K, which includes the memory needed for the operating system of the computer. Although the model is fairly large (238 endogenous and exogenous variables), only five quarters' worth of data are needed in core memory at any one time. Data for five quarters are needed because the model has lagged values of up to four quarters. After the model is solved for a given quarter (the fifth quarter in memory), the results for that quarter are written to the disk, a new quarter's worth of data are read (with the old first-quarter's data dropped from memory),

and the model is solved for the new quarter. This process is repeated throughout the prediction period. Because of this structure, the program does a fair amount of reading from and writing to the disk.

The program runs on IBM and IBM-compatible personal computers; it also runs on the DEC Rainbow and the Wang. It can be used on computers with only one disk drive, although two disk drives are more convenient. A version of the software allows one to use Lotus 123 to analyze the output; for this version the memory requirement is 192K. It takes about 10 to 20 seconds on a standard IBM PC to solve the model for one quarter. A twelve-quarter solution thus takes about 2 to 4 minutes.

## II. The Software

The software is menu driven and is easy to use even for students who have no prior experience using PCs. One of the main things that users must learn is the treatment of data sets. Included with the software is a "base" data set. This data set contains data on the endogenous and exogenous variables for the period of interest and all the other information, such as the coefficient estimates, that are needed to solve the model. For many applications the period of interest is an actual forecast period, for example, 1984:4 to 1987:4. In this case the values of the endogenous variables in the base data set are predicted values. If the period of interest is some historical period, the values of the endogenous variables in the base data set are the actual values. The teacher or student can use the "historical" part of the software to create a base data set for any six-year period from 1952 on.

Assume that the base data set is for an actual forecast period, so that the endogenous variable values are predicted values. This means that if a forecast is run with no changes made to the base data set, the pre-

\* Wellesley College, Wellesley, MA 02181, and Cowles Foundation, Yale University, New Haven, CT 06520, respectively. The software discussed herein is marketed under the name of FAIRMODEL by Economica, 2067 Massachusetts Avenue, Cambridge, MA 02140. The accompanying workbook is by Anthony Blackburn and Case (1985).

dicted values of the endogenous variables will simply be the values already in the data set. The job of the student is to change some of the information in the base data set and run a new forecast. The main menu that the user faces is as follows:

Enter	To
1	Change assumptions about monetary policy.
2	Change exogenous federal government fiscal policy variables.
3	Change exogenous state and local government variables.
4	Change exogenous foreign sector variables.
5	Change other exogenous variables.
6	Change equations by the use of add factors.
7	Drop or add equations.
8	Take equations to begin after the beginning of the forecast period.
9	Change coefficient values.
10	Exit.

The monetary and fiscal policy options (1 and 2) are the ones most often used. These are discussed in Sections III and IV below. Options 3, 4, and 5 allow other exogenous variables to be changed. Option 6 allows add factors to be used, something which is not of much interest for teaching purposes. Option 7 allows equations to be dropped, which means that the equations are not used and the endogenous variables that are explained by the equations are taken to be exogenous. This option is sometimes useful for examining subsets of the model's equations. Option 8 allows the equations to be dropped for only part of the forecast period. Finally, option 9 allows the coefficient values to be changed. An example of the use of this option is presented in Section III.

Once the changes have been made, a new forecast is run. New data are created from this run, which consist of the user's changes from the base data set and the new predicted values. The student can then use the software to examine the differences between the predicted values in the two data sets. These differences are the effects that the changes

have had on the endogenous variables in the model. As many new data sets can be made as the user likes, any previously created data set can be used as a base data set, and any two previously created data sets can be compared. The fact that all the information about an experiment is contained in one data set makes it very easy to compare alternative runs.

If the period of interest is historical (any period except the current forecast period), then the endogenous variable values in the base data set are the actual values, and a new "base" data set must be created before any changes are made. This data set is created by simply solving the model over the period of interest with no changes made. The endogenous variable values in this data set are then predicted by the model and are the appropriate ones from which to make comparisons when changes are made in running experiments. If the changes were made from the original base data set, the differences between the endogenous variable values in the new and base data sets would have included the predictive errors, which are not, of course, a result of the user's changes.

### III. Intermediate and Advanced Teaching

Knowledge of the structure of the model is needed for intermediate and advanced teaching. The theory behind the model is that agents make decisions by solving multiperiod optimization problems. The key decision variables for households are consumption and leisure, and the key decision variables for firms are prices, wages, production, investment, and employment demand. A typical estimated equation in the model has a decision variable on the left-hand side and variables that are assumed to affect the decision on the right-hand side.

The first step in teaching with the model is simply to go through the 30 estimated equations. There are 9 estimated equations for the household sector: three categories of consumption (service, nondurable, and durable), housing investment, four categories of labor supply (labor force participation of males 25-54, females 25-54, and all others, and the number of people holding two jobs), and



a demand for money equation. There are 12 equations for the firm sector. The 7 most important explain the firm sector's price level, production, investment, demand for workers, demand for hours per worker, the wage rate, and the demand for money. The other 5 equations explain overtime hours, dividends, interest, inventory valuation adjustment, and depreciation. The explanatory variables in the main equations are consistent with the view that households maximize utility and firms maximize profits. The equations are fairly easy to explain since one can appeal to students' knowledge of microeconomics.

The short-term interest rate in the model is explained by an interest rate reaction function of the Federal Reserve, and this is a good equation to discuss next. It is a "leaning against the wind" equation in the sense that the Fed is estimated to allow the short-term interest rate to rise as 1) inflation rises, 2) labor markets get tighter, 3) real output growth increases, and 4) lagged growth of the money supply increases. The inclusion of this equation in the model means that monetary policy is endogenous.

The remaining 8 estimated equations in the model include two term structure equations (explaining a long-term bond rate and a mortgage rate), a stock price equation, a demand for currency equation, an equation explaining bank borrowing from the Fed, a demand for imports equation, an equation explaining unemployment insurance benefits, and an equation explaining the interest payments of the federal government.

The next task is to review the 98 identities. A key feature of the model is that all flow-of-funds and balance-sheet constraints are met, and many of the identities relate to these constraints. There are six sectors in the model (household, firm, financial, foreign, state and local government, and federal government), and for each sector there is an identity determining its level of savings. The sum of savings across sectors is zero, since the revenue of one sector is the expense of some other sector or sectors. There is also an identity for each sector that relates the level of savings of the sector to changes in its assets and liabilities: any nonzero value of savings in a period must result in the change

in at least one asset or liability. Explaining these identities carefully for one sector allows one to go through them quickly for the other sectors. The remaining identities in the model, which are equations like the *GNP* definition, can also be quickly reviewed.

One is now in a position to have the students run some experiments. A good starting point is to have the change be an increase in federal government purchases of goods, which is a standard expansionary fiscal-policy action. This change can be analyzed under four different assumptions about monetary policy. In other words, four solutions of the model can be made for the fiscal-policy change, each solution corresponding to a different assumption about monetary policy. The four assumptions are: 1) the interest rate reaction function used (monetary policy endogenous), 2) the short-term interest rate unchanged from the base case, 3) the money supply unchanged from the base case, and 4) nonborrowed reserves unchanged from the base case.

Having the students explain carefully the reasons for the differences across the monetary-policy assumptions is an effective way of having them learn about the links between monetary policy and fiscal policy. Under each of the assumptions, the amount of government securities outstanding (the open market operations variable of the Fed) is endogenous, and the students should be required to explain why the government securities variable changed in the particular way that it did for each run. An example of a difference between runs is that the economy is more expansionary under the assumption of an unchanged interest rate than it is under the use of the interest rate reaction function. With the interest rate reaction function, the Fed responds to the fiscal-policy action by allowing interest rates to rise, which slows the expansion. The increase in government securities outstanding is larger in this case.

Explaining the four cases generally takes two or three class periods. There are two types of things to be learned from these results. One is the response of the household and firm sectors to the fiscal-policy change, and the other is the relationship between monetary policy and fiscal policy. The results

for the household and firm sectors can be related to the prior discussion of the individual equations for these sectors. The model has the feature that both an expansionary fiscal policy and an expansionary monetary policy stimulate private demand, and so to some extent one can consider the question of the relationship between monetary policy and fiscal policy without knowing too much about the detailed responses of the household and firm sectors. This makes the results somewhat less model specific than one might otherwise think. Also, accounting for the flow-of-funds and balance-sheet constraints is not model specific in the sense that these are just identities, and so anything that is learned from this accounting framework is not model specific. For example, crowding out issues can be easily examined since any deficit that the government runs from an expansionary fiscal-policy action must be financed by an increase in savings of at least one other sector, and the model tells one directly which sectors are doing the increased saving.

The next step is to run other fiscal-policy experiments. Various tax rates can be changed, the level of transfer payments can be changed, and the number of government jobs can be changed. These runs allow the effects of different fiscal policies to be compared (given the same monetary-policy assumption for each experiment). For example, a decrease in the personal income tax rate increases labor force participation in the model (a positive labor supply response), which, other things being equal, leads to a rise in the unemployment rate. The relationship between output changes and unemployment rate changes is thus different for this experiment than it is for the experiment in which government spending on goods is increased, where there is no direct labor supply response.

Pure monetary-policy experiments can also be run. The short-term interest rate, the money supply, or the level of nonborrowed reserves can be made exogenous and changed.

An example of an interesting advanced experiment is the following. First, run a particular experiment and record the results. Second, change the coefficients in the demand for money equations to alter the inter-

est sensitivity of the demand for money. Then run the same experiment for the different coefficient values, record the effects, and explain the differences. (This experiment will require a different base run because of the different coefficients used.)

Many other exogenous variables in the model can be changed. One of considerable current attention is the price of imports. An increase in this variable generally has the effect of increasing domestic inflation and decreasing real output. This negative correlation between inflation and output is contrary to the positive correlation that generally results when demand is stimulated by monetary or fiscal policy. These different correlations show the student that there is no simple relationship between inflation and output in the economy.

After a few weeks of going through results, students begin to understand simultaneous effects. They can think through a rather complicated series of variables affecting one another. They also learn that most changes take time to work themselves out in the economy.

As a final note, we have found that students often think of interesting experiments. One example is the following. The two term structure equations in the model are not consistent with the expectations theory of the term structure if expectations of the short-term interest rate are rational in the Muth sense. In other words, the model does not have rational expectations in the bond market. One student, however, forced the model to have this feature by dropping the two estimated term structure equations and adjusting the long-term rates to be consistent with the model's future predictions of the short-term rate. This was done by solving the model many times, each solution corresponding to a particular set of values of the long term rates, until, given the model's predictions of the short-term rate, the consistency requirement was met.

#### IV. Elementary Teaching

Elementary students are introduced to the basic concepts of macroeconomics through simple comparative static models. Textbooks begin with an explanation of "equilibrium

national income determination." The simple aggregate supply/demand equilibrium model continues to be a very useful pedagogical tool. It provides a basic structure within which theoretical debates and policy options can be discussed. Although many concepts and alternative theories can be incorporated into this structure (for example, theories of consumption/savings behavior), there are a number of problems that result from its use.

One problem is that students often come away from principles courses with a very inaccurate impression of time lags and of the magnitude of potential policy impacts. Shifting the aggregate demand curve in a static model, for example, will reveal the direction that most variables will move, but not the magnitude or timing of the movements. Good students in principles courses will be able to tell you that a personal tax cut is likely to increase disposable income, initially increase both savings and consumption, and ultimately raise the "equilibrium level of *GNP*" by a multiple of the original tax cut. What they cannot tell you is how long it takes to get to the new "equilibrium," or how large the ultimate change in *GNP* is likely to be.

In teaching the multiplier we speak of "periods" (as smaller and smaller disequilibria lead to the new equilibrium), but the periods are used only to teach the character of the adjustment process and they are never tied to actual intervals of time. The multiplier values that students think exist in the real world are usually not sensible. It is not uncommon for even good students to think that the multiplier is greater than five "since the marginal propensity to save is less than .2."

The use of the software for the Fair model provides a useful way of teaching students about the adjustment process and about the likely size of the policy effects. The model can be used for this purpose without having to teach the detailed specification of the model. For example, students can simply take the base forecast for some period, increase real government purchases of goods by one billion dollars, and observe and plot the resulting changes in real *GNP* over the period. This can be done under the various assumptions about monetary policy, and by running these experiments the students can

get a good idea of both the size of the policy effects over time and the influence of monetary policy on fiscal policy.

Another important use of the software at the elementary level is examination of the historical data. Elementary students receive only a cursory exposure to the empirical side of macroeconomics and recent economic history. If data are discussed at all, they are presented in simple tables. Elementary texts do contain descriptions of recent events, but students "hands on" learning is generally accomplished by working through exercise sets built around simple hypothetical economies ("the king decides to stimulate demand with a tax cut of 10").

Experiments with the Fair model use real data and can be constructed around actual historical events. To prepare for the experiments, it is useful to have students do a series of descriptive exercises with the quarterly macroeconomic data. Over the last fourteen years, when was the unemployment rate the highest? What happened to real *GNP* over the same period? Explain the relationship between real and nominal *GNP*. How much did real *GNP* per capita grow between 1960 and 1970? Between 1970 and 1980? Describe the behavior of the money supply between 1970 and 1975.

Once the students have some familiarity with the data, the simulation experiments can begin. The workbook that accompanies the software contains 49 experiments. Most experiments begin with a newspaper or magazine article describing some event, controversy, or historical period. For example, last year there was a widely publicized debate between Martin Feldstein and Donald Regan about whether the Fed should hold the line at 6 percent money growth or move to an 8 percent target. The model can be used to simulate these two policies, and students can compare the results. Which yields higher interest rates? Is there an impact on nominal *GNP*? On real *GNP*? Explain the links. What happens to projected inflation under the two policies?

Another experiment is more historical. President Ford asked for and got a substantial tax cut in the spring of 1975 (newspaper article). At the same time the Fed was pursuing an easy monetary policy (look at *M1*

growth and interest rates). Both policies were in response to the recession that hit bottom in May of 1975 (verify). A modest "recovery" began in the second half of 1975 (what is meant by recovery?). Now go back to 1975 and do away with the tax cuts. What would have happened? Now restore the tax cuts, but hold money growth to 4 percent. Who was more responsible for the recovery, Arthur Burns or Gerald Ford, or is this question meaningful?

Other experiments include an examination of the effects of policy changes on labor supply and investment. In 1981 Robert Eisner did a major project for the Office of Tax Analysis at the Treasury Department in which he simulated the corporate tax cuts of the 1981 Economic Recovery Tax Act using

several large econometric models (DRI, Chase, etc.). Nearly identical experiments were run by 37 undergraduate Public Finance students at Wellesley College who analyzed the supply-side impacts of smaller and larger tax cuts.

## REFERENCES

- Blackburn, Anthony and Case, Karl, *FAIR-MODEL Student Manual: An Economic Laboratory in Theory, Policy and Forecasting*, Englewood Cliffs: Prentice Hall, 1985.
- Fair, Ray C., *Specification, Estimation, and Analysis of Macroeconometric Models*, Cambridge: Harvard University Press, 1984.

# Cost Effectiveness of Computer-Assisted Economics Instruction

By DARRELL R. LEWIS, BRUCE R. DALGAARD, AND CAROL M. BOYER\*

Is computer-assisted instruction (CAI) in economics cost effective? The early enthusiasm and verified instructional effectiveness of CAI during the 1960's and early 1970's led many to believe that CAI adaptations would increase into the 1980's. Although CAI was especially visible in the teaching of college-level economics, a recent review of the literature indicates that the use of CAI in economic education appears to have peaked during the early to mid-1970's (Dalgaard et al., 1984). The decline that followed is often attributed to questions of the cost feasibility and cost effectiveness of CAI in both economic education (John Soper, 1974) and in education generally (Robert Butman, 1973; Denyse Forman, 1982; and Greg Kearsley, 1977). In light of recent advances in computer technology and recent marketing innovations in the development, dissemination, and use of mini and micro computers (including concurrent cost reductions), many of the previously discontinued CAI activities may now be both cost feasible and cost effective, and should therefore be reconsidered.

Unfortunately, however, systematic efforts to collect data on the costs of CAI are seriously lacking. Those few studies that do report cost data do so only for mainframe-type CAI systems that, for the most part, are now at least five to ten years out of date. Clearly what is needed today is not only more current, carefully designed cost information, but also systematic assessments of the cost feasibility and cost effectiveness of both existing and new CAI systems.

## I. Is CAI Effective?

Within the field of economic education, surprisingly little is known about the effectiveness of CAI. For example, virtually nothing has been reported in the literature on the use of CAI in elementary and secondary schools. Although one could generalize from studies that focus on other subject matter areas, additional studies clearly are needed in economic education at these pre-college levels of instruction.

Even at the college level, only a few studies have systematically examined the effectiveness of CAI in economic education. Of the twenty-one studies of college-level CAI reviewed by Soper, for example, only six had included any empirically based assessment of student performance (i.e., achievement and attitudes); fewer still had made any reference to costs. A more recent review by John Siegfried and Rendigs Fels (1979) focused on seven experiments that had assessed college-level CAI in terms of its impact on cognitive achievement, student attitudes, retention, and the distribution of benefits between high- and low-achieving students. All seven reported positively on the effectiveness of CAI. Only two other carefully designed studies of the effectiveness of CAI in economics are reported in the literature (Donald Paden and Michael Barr, 1980; James Marlin and James Niss, 1982).

Unlike the recent literature in economic education, the education research literature has shown a consistent interest and involvement in CAI at all levels of instruction. Other reviewers generally support the effectiveness of CAI and usually conclude that it is effective in improving student achievement. At the elementary school level, for example, a number of extensive reviews (see especially Forman; Dean Jamison et al.,

\*Professor, Associate Professor, and Research Specialist, respectively, University of Minnesota, Minneapolis, MN 55455-0211.

1976; James Kulik et al., 1983) all reported results from a large number of independent studies showing substantial instructional advantages for CAI. Although it has proven more difficult to show the achievement advantage of CAI at higher levels of instruction, these same reviewers generally concluded that CAI was at least as effective as conventional instruction at the secondary and college levels.

Perhaps the most important and consistent finding in the education research literature concerning the effectiveness of CAI is its effect on instructional time. Although seldom addressed in most studies, whenever instructional time has been examined, CAI has been found to have profound effects in reducing instructional time for students. In studies from both school and industry settings, CAI has been reported as saving anywhere from 30 to 90 percent of student learning time (Butman; Kulik et al., 1983). Paden and Barr, for example, reported that CAI students not only performed better in economics than non-CAI students, but that they also spent less time studying than their non-CAI peers.

When student attitudes toward the subject matter and the quality of instruction are assessed, they are generally favorable with respect to CAI (in economic education, see especially Bernard Booms and D. Lynne Kaltreider, 1974; Paden and Barr). Other reviewers also have found that the effect of CAI is even stronger and more positive on student attitudes toward computers than on attitudes toward the subject matter and the quality of instruction.

In addition to the positive effects that CAI reportedly has on student achievement, attitudes, and instructional time, the research literature reveals several other advantages for CAI, including its capability for individualizing the instruction process; for keeping students informed of their progress through immediate feedback and achievement summaries; for simulating experiences not possible without a computer; and for allowing students to review previous instruction, request special help, or continue on to extra learning activities (in economic education,

see especially Allen Kelley, 1972; William Davisson and Frank Bonello, 1976).

## II. Is CAI Economical?

If the above generalizations are valid, then one should expect the use of CAI in some instructional settings to contribute to enhanced efficiency. At the least, in assessing alternative methods of instruction, one must examine all of the likely costs and outcomes. Clearly, it is necessary to establish a basis for comparing CAI with alternative methods of instruction in terms of costs and outcomes before any assessment can be made of the alleged efficiency or cost effectiveness of CAI. Most importantly, even when no significant differences in student achievement are realized from the introduction of CAI into an instructional setting, one should expect to realize (and measure) changes both in *student-teacher ratios* and *student instructional time*.

To illustrate, consider the following example. Assume an instructional setting where students with CAI score about the same as students with conventional instruction. Classroom instructor costs are assumed to be \$1.40 per student hour. With CAI, however, we note that each instructor can manage (i.e., teach) twice as many students. Consequently, instructor costs per student hour can be reduced by one-half or \$0.70. Add \$1.50 for the terminal-hour cost of a CAI system (computed as the average total cost per hour) and \$3.60 per hour for student time (opportunity cost for lost wages) for a total cost of \$5.80 per student hour on the computer system. Average time-improvement factors for CAI of 0.7 are not unreasonable (Butman), so final total cost per equivalent classroom hour is about \$4.00. This might compare with a cost of \$1.40 per hour for the instructor plus \$3.60 for student time or \$5.00 total for conventional instruction. Given the above assumptions, a cost savings of approximately \$1.00 per hour is forecast for CAI in this illustration. In other words, CAI is 20 percent more efficient than conventional instruction.

As a real example, when Deltak, Inc. (1981)

compared its industrial training programs and considered the cost of student time, it found that a five-day, instructor-led course of 10 students was more costly than a computer-enhanced, learner-paced, multimedia approach at a ratio of \$1,120 to \$680. In short, in this example, conventional classroom training was 65 percent more expensive than CAI.

Computer-assisted instruction can contribute to instructional efficiency (and can be more cost effective than conventional instruction) *provided* certain instructional outcomes are realized. In simplest form, CAI can be cost effective if (a) it costs the same as conventional instruction but contributes more to student achievement in the same amount of instructional time, or (b) it costs less or results in less training/instructional time for students to achieve the same learning outcomes as conventional instruction, and/or (c) it results in an increase in student-teacher ratios (which can be translated into lower overall costs to the system).<sup>1</sup>

### III. Cost Analysis of CAI

It should be clear from the above discussion that one needs cost information in order to make fully rational decisions about CAI. Unfortunately, very little hard data are reported in the literature. Often what information does exist has limited application. For example, almost all of the early investigations of cost factors related to the use of CAI explored only their cost feasibility. Specifically, these studies accumulated cost data related only to the acquisition, implementation, and maintenance of CAI. It is important to note that cost-feasibility questions do not consider student performance (i.e., achievement and attitudes) relative to the

costs of alternative methods of instruction. Costs may be compared but not with student performance in mind. Moreover, studies that do address the cost feasibility of CAI are really addressing budgetary costs and not economic costs.

Cost-effectiveness studies, on the other hand, consider some objective measure of student performance in alternative instructional settings, with the cost of delivering instruction as the determinant. Almost all of the studies identified in the previous section related only to some measurement of the effectiveness of CAI as compared to conventional instruction. Very few studies reported cost data in any form. Although a few cost-feasibility studies of CAI are emerging in the literature, cost-effectiveness studies of CAI are still noticeable by their absence.<sup>2</sup>

There is some precedent for the application of cost-effectiveness analysis to instructional delivery systems. Martin Carnoy (1976) and J. Mayo et al. (1975), for example, analyzed the economic costs and benefits of educational television; and K. Lumsden and C. Ritchie (1975) attempted to analyze the cost effectiveness of educational technology in a university setting. Few studies, however, have examined issues relating to costs or cost effectiveness of CAI. To illustrate, we can find virtually nothing on costs in the 59 studies of college-level CAI evaluations that were examined by Kulik et al. (1980). The 59 studies reported only on student achievement, the correlation between student aptitude and achievement, course-completion rates, student attitudes, and instructional time. A meta-analysis by Kulik and his colleagues indicated that 54 of the 59 studies reported on student achievement whereas only 8 of the 59 studies reported on instructional time. The latter is the *only* variable in

<sup>1</sup>Unfortunately, most of the current use of CAI has been in the form of additions to conventional instruction and thus precludes realizing many of these cost savings. This has undoubtedly resulted from the reality that most school and classroom settings do not permit the instructional and curricular flexibility necessary to substitute capital (in the form of new technology such as CAI) for labor in the teaching/learning process.

<sup>2</sup>This paper is a partial product of a larger project wherein we are developing a more formalized cost function with parametric estimations derived from field-based cost data on the use of CAI with micro computers in the public schools. We expect to relate such cost information to various forms of outcome data in order to estimate cost effectiveness for alternative instructional settings.

most of the reported studies that might allow for any cost analysis. In a subsequent study by Kulik et al. (1983), a meta-analysis of 51 studies on CAI at the secondary level yielded similar results. Again, no mention was made of the costs of CAI except in two studies that reported only on instructional time.

It is both surprising and disappointing to note that out of literally hundreds of citations in the education research literature on CAI, only six studies have reported any systematic examination of the empirical basis of costs (Fred Hofstetter, 1983; Jamison et al., 1970; Jamison et al. 1976; Kearsley; Henry Levin and Louis Woo, 1981; Paden and Barr). Moreover, only the studies by Jamison et al. (1970) and Paden and Barr systematically compared CAI costs in terms of value-added in student achievement with those of conventional instruction and reported on the relative success and cost effectiveness of CAI. An earlier study by Kelley did report in summary terms on the perceived cost effectiveness of TIPS (Teaching Information Processing System) in economic education, but unfortunately no empirically based cost data were presented or analyzed.

In summary, few studies in the education literature address the cost factors associated with CAI, even fewer systematically consider student instructional time and the cost effectiveness of CAI, and none addresses the cost factors associated with the use of micro computers in CAI.

#### IV. Future Directions for CAI

A recent national survey by Jack Chambers and Alfred Bork (1980) assessed the current and projected use of the computer in U.S. public secondary and elementary schools, with special emphasis on the use of new micro CAI systems. The results showed that 74 percent of all districts responding to their survey used the computer to support their instructional process and 54 percent used CAI. Furthermore, 75 percent of the districts indicated that their teachers were interested in developing new CAI programs. Chambers and Bork concluded that in the next decade, "cost reductions due to mass production and consumption for home enter-

tainment and learning will permit cost-effective uses of computer-assisted learning in both the traditional classroom and in other settings" (p. 5).

Given these current and projected activities for CAI and the micro computer in the U.S. public elementary and secondary schools, why has so little been reported, especially in the recent research literature in economic education? We believe the answer lies both in the research literature and in a number of institutional circumstances within higher education.

First, it is important to note that until recently almost all software and program developments for CAI have taken place in the offices of university faculty. Moreover, and again until only recently, most faculty were familiar and worked almost exclusively on mainframe, research-type computer systems. At the same time, the central computer facilities of most universities were expensive, perceived as fixed costs in the short run, and selected and purchased for their research capabilities. Consequently, both faculty and central administrators have had strong incentives to preserve the functional form of existing computer systems. Therefore, many universities have been more than willing to provide departments with CAI services for only their marginal costs, and often even free, simply because such provision of service provides additional rationale for use by central administration with regents and legislators for the preservation, or even expansion, of their present computer systems.

Second, given this focus on mainframe computer systems, it appears that by the mid-1970's many users of CAI—at least in college-level economics—were becoming disillusioned about its likely cost effectiveness (Soper). As evidenced by the lack of attention to CAI in the economic education literature, its rate of use even at the college level appears to have dropped off appreciably. It also is important to note that in many cases when external grants and departmental subsidies were withdrawn (as, for example, in the cases of two rather large-scale CAI projects in economics at Illinois and Wisconsin), the projects were often similarly terminated or drastically reduced. The realities of equip-



ment and on-line costs in the mid-1970's apparently dictated a lack of cost effectiveness for CAI at the college level. Even today, after ten years and millions of dollars spent on development and marketing costs by Control Data Corporation, only seven universities in the United States have adopted some form of the PLATO mainframe system for teaching economics. At least in the past, the major reason that many such projects were considered not to be cost effective was that the hardware and delivery systems were so expensive. Although there were undoubtedly other cost savings (such as trading capital for labor, freeing instructional staff for other educational services, and saving student instructional time), the perceived costs of CAI technology apparently precluded university-funded continuation of the projects at that time.

In short, the biases and fixed costs within higher education have precluded much flexibility, many resources, and much attention to the use of micro computers in CAI. Moreover, little in the literature concerning the costs of CAI has been either reported or persuasive.<sup>3</sup> At the same time, however, the public schools have had few such hindrances or disincentives, and the private companies have been actively pursuing these targets of opportunity.

As discussed above, the research literature dealing with CAI does provide evidence that CAI can be effective in the teaching/learning process within some instructional settings. Nonetheless, the fundamental question remains, is CAI *currently* cost effective?

The literature, at best, only gives us very limited information on this critical question. For most of the reported uses of CAI and at

all levels of instruction, we still do not know the answer to the question of whether CAI is cost effective. Our review of CAI cost analysis and evaluation studies in both economic education and education generally indicates that, at least at the college level through the mid-1970's when almost all CAI was on mainframe type systems, CAI probably was *not* cost effective. On the other hand, with the current costs of computer hardware less than 20 percent of what they were only seven years ago, and with the costs of such hardware and communication systems making up approximately 50 to 90 percent of the total costs for CAI (Dalgard et al.), it is safe to assume that the current costs of using micro computers are probably *less than 30 to 60 percent* of the total costs at the time the earlier studies were conducted. It also is safe to assume that any reduction in the costs of technology has been accompanied by increases in labor costs. Together, these factors make CAI a strong candidate today for reexamination and possible adoption at all levels of instruction. In short, CAI may have become both cost feasible *and* cost effective!

## REFERENCES

- Booms, Bernard H. and Kaltreider, D. Lynne, "Computer-Aided Instruction for Large Elementary Courses," *American Economic Review Proceedings*, May 1974, 64, 408-13.
- Butman, Robert C., "CAI—There Is a Way to Make It Pay (But Not in Conventional Schooling)," *Educational Technology*, December 1973, 13, 5-9.
- Carnoy, Martin, "The Economic Costs and Returns to Educational Television," in Gene V. Glass, ed., *Evaluation Studies: Annual Review*, Beverly Hills: Sage Publications, 1976.
- Chambers, Jack A., and Bork, Alfred, *Computer Assisted Learning in U.S. Secondary/Elementary Schools*, Report 80-03, Center for Information Processing, California State University-Fresno, July 1980.
- Dalgard, Bruce, Lewis, Darrell R. and Boyer, Carol M., "Cost and Effectiveness Considerations in the Use of Computer Assisted Instruction in Economics," *Journal of Eco-*

<sup>3</sup>It is important to note that, in addition to the lack of cost information on the new micro computer technology, a second factor limiting the implementation of CAI via micro computers is likely to be the unavailability of inexpensive software. In this regard, a number of commercial companies are rapidly developing such software, especially for use in the elementary and secondary schools. The Joint Council on Economic Education (1984), through its recent National Computer Economic Education Program, also is targeting CAI as an area of needed curriculum development and is currently supporting software development efforts.

- nomic Education*, Fall 1984, 15, 309-23.
- Davisson, William I. and Bonello, Frank J., *Computer-Assisted Instruction in Economic Education*, Notre Dame: University of Notre Dame Press, 1976.
- Forman, Denyse, "Search of the Literature," *The Computing Teacher*, January 1982, 5, 37-50.
- Hofstetter, Fred T., "The Cost of PLATO in a University Environment," *Journal of Computer-Based Instruction*, Spring 1983, 9, 148-55.
- Jamison et al., Dean T., "Cost and Performance of Computer-Assisted Instruction for Education of Disadvantaged Children," in J. Froomkin et al., eds., *Education as an Industry*, Cambridge: Ballinger, 1976.
- Jamison, D., Suppes, P. and Butler, C., "Estimated Costs of Computer Assisted Instruction for Compensatory Education in Urban Areas," *Educational Technology*, September 1970, 10, 49-57.
- Kearsley, Greg P., "The Cost of CAI: A Matter of Assumption," *AEDS Journal*, Summer 1977, 10, 100-12.
- Kelley, Allen C., "TIPS and Technical Change in Classroom Instruction," *American Economic Review Proceedings*, May 1972, 62, 422-28.
- Kulik, James A., Kulik, Chen-Lin C. and Cohen Peter A., "Effectiveness of Computer-based College Teaching: A Meta-analysis of Findings," *Review of Educational Research*, Winter 1980, 50, 525-44.
- \_\_\_\_\_, Bangert, Robert L. and Williams, George W., "Effects of Computer-Based Teaching on Secondary School Students," *Journal of Educational Psychology*, 1983, 75, 19-26.
- Levin, Henry M. and Woo, Louis, "An Evaluation of the Costs of Computer-Assisted Instruction," *Economics of Education Review*, Winter 1981, 1, 1-25.
- Lumsden, K. and Ritchie, C., "The Open University: A Survey and Economic Analysis," *Instructional Science*, 1975, 4, 237-92.
- Marlin, James W. and Niss, James F., "The Advanced Learning System, A Computer Managed, Self-paced System of Instruction: An Application in Principles of Economics," *Journal of Economic Education*, Summer 1982, 13, 26-39.
- Mayo, J., McAnany, E. and Klees, S., "The Mexican Telesecundaria: A Cost-Effectiveness Analysis," *Instructional Science*, 1975, 4, 197-236.
- Paden, Donald W. and Barr, Michael D., "Computer-Assisted Instruction in an Elementary College Economics Course," *Computers and Education*, 1980, 4, 259-67.
- Siegfried, John J. and Fels, Rendigs, "Research on Teaching College Economics: A Survey," *Journal of Economic Literature*, September 1979, 17, 923-69.
- Soper, John C., "Computer-Assisted Instruction in Economics: A Survey," *Journal of Economic Education*, Fall 1974, 6, 5-28.
- Deltak, Inc., Pamphlet, Oak Brook, IL, 1981.
- Joint Council on Economic Education, "Joint Council Developing Computer Software in Economics," *Update*, Winter 1984, 1, 5.

# THE DEREGULATION OF BANKING IN THE UNITED STATES<sup>†</sup>

## Legislative Construction of the Monetary Control Act of 1980

By RICHARD H. TIMBERLAKE, JR.\*

[The Monetary Control Act] is one of the most complex and least understood pieces of legislation that I have ever seen come before a legislative body. It has been referred to as everything from a Christmas tree to a forest primeval, the latter probably being the more appropriate phrase.

U.S. Senator Donald Stewart,  
*Congressional Record*,  
March 27, 1980, p. 6910

The Depository Institutions Deregulation and Monetary Control Act (DIDMCA) of 1980 was hailed by Senator William Proxmire, one of its sponsors in the U.S. Senate, as the "most significant banking legislation since [the passage of the Federal Reserve Act in] 1913." Title I in the bill extended the Federal Reserve's control over reserves to all depository institutions and added some other powers as well. Title II called for the phasing out of all government-imposed interest rate ceilings on banks and other financial institutions and generally removed competitive inequities between them. The Act has in total eight Titles, which deal with Truth in Lending Simplification, State Usury Laws, Amendments to the National Banking Laws, and other matters. The first two Titles, however, contain the important substance of the Act as well as its contradictory implications: The power of the Fed and its regulatory scope are greatly extended. At the same time, restrictions on freedom of economic activity

for the rest of the banking and financial system are significantly relaxed.

Economists and bankers have concurred with Senator Proxmire on the importance of the DIDMCA of 1980. They have long recognized the desirability of regulatory simplification, such as homogeneous reserve requirements for banks and abolition of interest rate ceilings on time deposits, so they have applauded the deregulatory aspects of the Act. They have been remiss, however, in failing to point out the disparate treatment of institutions implied by the two major "Titles," and especially the significance of the additional powers granted the Fed.

The debates on the bill in both Houses of Congress reveal the procedure by which these two dissimilar provisions were incorporated into the same law. Fundamental to this result were the opinions and testimony of Federal Reserve Board spokesmen who significantly influenced congressional committees and were thereby instrumental in gaining additional powers for the Fed.

### I. The Ancestry of the DIDMCA of 1980

The Monetary Control Act of 1980 began in 1979 in the House of Representatives as the Consumer Checking Equity Act of 1979 (H.R. 3864) and the Monetary Control Act of 1979 (H.R. 7). Also proposed earlier were three different Senate bills: S. 85 would have made membership in the Federal Reserve System mandatory for all depository institutions, S. 353 would have had the Fed pay market interest rates on the reserve account balances of Federal Reserve member banks, and S. 1347, another Monetary Control Act of 1979. H.R. 7 was the only one of these earlier bills to pass either House before the final bill. It passed the House of Representatives in 1979, but its counterpart in the

<sup>†</sup>*Discussants:* William Beranek, University of Georgia; Richard K. Vedder, Ohio University.

\*Department of Banking and Finance, College of Business, University of Georgia, Athens, GA 30602. References to the *Federal Reserve Bulletin* are abbreviated by *FRB*, and to the *Congressional Record* by *CR*. Comments by William Beranek, Alfred Bornemann, George Selgin, and Richard Vedder were very helpful.

Senate, S. 1347, was not debated. All of these earlier bills served as "ancestors" for the bill, H.R. 4986, that finally passed both Houses in 1980.

## II. The "Problem" of Fed Membership

Hearings and deliberations on these various House and Senate bills began in early 1979. On February 26 of that year, G. William Miller, who was then Chairman of the Fed's Board of Governors, commented at length before the Senate Committee on Banking, Housing and Urban Affairs on both S. 85, the "Proxmire bill," and S. 353, the "Tower bill."

The first issue Miller addressed was the attrition of Fed membership. Banks had been steadily leaving the Federal Reserve for over 25 years; between 1970 and 1978, the decline numbered about 300 banks (out of about 14,500). The decline in membership was accompanied by a reduction in the proportion of members' deposits to total bank deposits from 81 to 72 percent. The explanation for this decline was seen to be the high interest rates caused by inflation. Fed member banks were required to keep zero-interest reserve balances with the Fed Banks of their districts. They also were entitled to certain services; such as the clearing of checks and access to the discount window. As nominal interest rates rose behind inflation during the 1970's, these zero-return reserve accounts became more burdensome. The new freedom achieved by nonbank financial institutions that allowed them to create quasi-demand deposits and to pay interest on them required a similar allowance for banks so that the latter could stay competitive in the struggle to get and keep depositors (Miller, p. 230).

All Fed officials, who subsequently testified on the various bills Congress considered, emphasized this "problem." Their argument was that loss of membership also meant a loss of Federal Reserve control over the banking system. "All the legislative proposals" Paul Volcker stated in his testimony, "need to be judged first of all against the central objective: We need to strengthen our ability to implement monetary policy in a variety of possible circumstances..." (1979, p. 823).

The bills, H.R. 7 and S. 85, he noted, would be accompanied by substantial reduction in "reserve balances that would be held in Federal Reserve Banks. These balances," he asserted, "*and only these balances*, provide the 'fulcrum' for the efficient conduct of monetary policy" (1979, p. 825). Volcker's views were echoed a few months later in the House by Henry Reuss of Wisconsin, who stated: "What this bill does...is give the monetary authorities a power they desperately need [sic] if they are going to pursue [an anti-inflationary policy]... Unless they have a 'fulcrum,'... a reserve base upon which they can conduct their open market policy, they are incapable of regulating the money supply" (CR, 125, 1979, pt. 15, p. 19689).

The Fed's line of argument is open to the criticism of an invalid premise—the implication that Federal Reserve monetary control depends critically on member bank reserve accounts in Fed Banks. Left out of this accounting are the banks' holdings of vault cash—Federal Reserve notes—that increased from \$7 billion in December 1970 to \$17 billion by the end of 1978. These notes are also legal reserves. They, too, are supplied by the Federal Reserve System in the same manner as reserve accounts, and are held by both member and nonmember banks and by all other financial institutions.

These two items together compose the major elements of the monetary base, the total of which is under Fed control on a day-by-day basis no matter what the values are of each constituent part. As the base is changed by Fed policies, so are changed in varying degrees the components of the conventional money stocks,  $M_1$  and  $M_2$ . Furthermore, Fed authorities know accurately the quantity of Federal Reserve notes outstanding, how much are held by member banks, and how much by all banks. Even if Fed membership were to drop to zero and if no banks were required to hold reserves, any creation of "transaction accounts" would require depository institutions to hold some kind of providential reserves in the form of monetary base items to meet liquidity demands by other depository institutions and depositors. Since the Fed manufactures virtually all of the monetary base, which is in

turn required for the existence of any common money, Fed control is complete no matter what volume of reserve balances is held at Fed Banks.

The Fed solution to the membership problem was to encourage Congress to make reserve requirements mandatory for all depository institutions. H.R. 7, S.85, and the bill that finally passed both Houses of Congress, H.R. 4986, contained this provision. Institutions so affected could still choose to remain nonmembers of the Fed, but they would all have to keep zero-interest reserve accounts with Fed Banks. In addition, they would all be entitled to enjoy the privilege of the discount window.

This perquisite was emphasized in Fed testimony. Use of the discount window had shrunk, noted Miller, due to membership attrition. Nevertheless, the window was still "the lender of last resort" to the payments system. "If the proportion of institutions having access to this facility were to decline," he warned, "individual institutions...could not...cushion temporarily the impacts of restrictive monetary policies.... Thus, the Federal Reserve may find that its ability to limit growth in money and credit in order to curb inflation was being unduly impeded because the safety valve provided by the discount window was gradually losing its effective coverage" (p. 230).

Miller's final argument is absurd. The Fed's powers to promote inflation, as well as to abate inflation, have nothing to do with the discount window. Furthermore, his remarks state that the Fed can "limit growth in money and credit in order to curb inflation" without a discount window and without direct control over bank reserves.

In any case, the contemporary Fed has powers far beyond the scope of discount window accommodation. It creates monetary base constantly and positively, not just when an occasional bank needs extra liquidity. To put it baldly, virtually every hour of the day the Fed is acting as a "lender of last resort" by creating notes and reserves via open market purchases of government securities. In this holistic creation of money, the role of the discount window is negligible. The monetary base at the end of 1980 was \$155 billion

while loans to member banks at this time were less than \$1.5 billion. As a source of monetary base material, discounts to member banks were therefore less than 1 percent of the total.

Not only has discounting become a negligible sideline of Fed policy in the creation of monetary base, but banks have a much broader area in which to market their reserve needs or bounties—the Federal funds market. The activity in this market swamps the minor operations of the discount window. Fed funds and repurchase agreements just for large member banks were around \$108 billion in late 1980, compared to less than \$1.5 billion in Fed loans to members (*FRB*, 1981, p. A6).

No essential function in the payments system would be harmed in the slightest by the abolition of the discount window. It is an anachronism. Its only "function" is to act as a Federal Reserve showpiece—a public relations device to suggest to Congress, the media, and the general public that the Fed is some kind of an insurance agency against financial disaster.

### III. Federal Reserve Note Collateral

One of the more intriguing sections in Title I of the DIDMCA comes under the subsection labeled "Miscellaneous Amendments." It expands the "eligible collateral" provision for Federal Reserve notes outstanding to include the fully guaranteed obligations of a foreign government or the agency of a foreign government, as well as any other financial assets that may be purchased by Reserve Banks (*FRB*, 1980, p. 448). The question implied by this provision is, why would the Fed need more collateral for notes when it could already monetize unlimited quantities of government securities that were all eligible collateral?

The bill H.R. 7, when first put together in 1979, had no provision for any change in collateral allowances. Congressman Ron Paul noted that only in the third version of H.R. 7 did the collateral provision surface for the first time, "but...without one word of [formal] testimony from the Fed" (*CR*, 125, 1979, pt. 15, p. 19678).

The Fed had quietly worked in the amendment at the conference committee meetings (*CR*, 125, 1979, pt. 15, p. 19678). Congressman William Stanton of Ohio noted that the amendment had been "added to the bill at the last moment without any testimony from the Fed or any other witness" (*CR*, 125, 1979, pt. 15, p. 19679). The objections of Paul, Stanton, and others succeeded in getting the collateral amendment deleted. The bill then passed the House by a vote of 340-20 (*CR*, 125, 1979, pt. 15, p. 19690).

In the next session of Congress the bill (H.R. 4986) that ultimately became the DIDMCA of 1980 was introduced. It was debated in both Houses, and ended up in another House-Senate conference committee to resolve disagreements. The conference committee report then went back to each House for consideration.

Even though the final bill mandated universal reserve requirements controlled by the Fed, it also anticipated a general reduction of reserve requirements overall to a uniform value around 12 percent. Such a reduction implied that total reserve balances kept at Federal Reserve Banks, plus vault cash reserves in the form of coin and Federal Reserve notes, would probably decline (the Fed estimated) by about \$15 billion. This change was to take several years so that depository institutions would experience no immediate hardships or windfalls.

When Senator Proxmire presented the bill to the full Senate, the provision that anticipated reduction of reserve balances at the Fed had suddenly become a "problem" requiring the re-inclusion of the infamous collateral provision. The reason for Proxmire's "problem," however, resulted from his programmed misinterpretation of the Fed's procedures for creating money. "A portion of the Federal Reserve's securities portfolio," he explained, "...represent[s] purchases made [by the Fed] *with reserves deposited by member banks*. Since the Monetary Control Act would release [sic] about \$15 billion in reserves, a comparable amount of securities would need to be sold. This would reduce the collateral available for Federal Reserve notes [!]." Such a reduction, Proxmire con-

tinued, could proceed to the point where the Fed had inadequate collateral for the issue of notes. The supplementary collateral provisions in the bill would alleviate this "technical" problem (*CR*, 126, pt. 6, 1980, p. 6897, emphasis added).

In the House debate over H.R. 7 several months earlier, Chalmers P. Wylie of Ohio had similarly "explained" the Fed-commercial bank relationship to his House colleagues: "The banks hold some of these reserves [created by open market operations] as vault cash and the rest goes *into* the Federal Reserve System which it uses for investments—the return from which goes to the Treasury of the United States [sic]" (*CR*, 125, pt. 15, 1979, p. 19669, emphasis added). With these "explanations" of Fed operations, any resulting legislation could hardly be rational.

Volcker had earlier testified before Proxmire's committee and had cultured this twisted "explanation." Reductions in reserve requirements for depository institutions, he had stated, "could be technically unworkable because [they] might result in insufficient amounts of...eligible financial assets to meet the collateral requirements against notes." The House, Volcker observed, had rejected the extension of collateral provisions in the H.R. 7 bill. To meet these objections, he suggested adding to the present list "only assets acquired abroad arising from time to time out of our foreign currency operations." Specifically, he wanted "short-term foreign government securities" (1979, p. 828). In a later statement made on February 4, 1980, Volcker again mentioned the "problem," and volunteered that "...we are prepared to supply an appropriate amendment that could be attached to any bill that would deal with the problem" (p. 147).

The Conference Committee received the amendment and accepted Volcker's rationale for it. As everyone acquainted with Fed operations knows, the money-creating process works exactly opposite the way "explained" by Proxmire and Wylie. As the Fed buys the securities, it perforce creates member bank reserve accounts or Federal Reserve notes for which the government securities serve as

collateral. The Fed can never be short of collateral for its monetary base "obligations" because it has always created exactly the same dollar amount of monetary base as the dollar value of the securities it has purchased.

Regardless of collateral, Federal Reserve notes are "legal tender for all debts, public and private." Eligible collateral is at best simply a security provision that limits the items the Fed is allowed to monetize. It is a leftover from the time 50 to 70 years ago when the notes were not legal tender and the Fed was an entirely different kind of institution. Collateral security is a pointless flourish for legal tender currency.

Actual experience since the passage of the DIDMCA indicates that no problem would have arisen (as indeed it *could* not) with only conventional collateral available for Fed notes. Even though reserve accounts at Fed Banks did decline as reserve requirements were phased down toward 12 percent, the Fed's consolidated statement during the period 1981-84 never showed less than an excess of \$9 billion in conventional collateral over the outstanding issues of Federal Reserve notes (*FRB*, 1981-83, Tables 1.18, p. A-11).

After the DIDMCA of 1980 had been operational for a few years, Charles Partee of the Federal Reserve Board appeared again before the House Banking Committee and reported on the Fed's use of the new collateral provision. He noted that Fed holdings of foreign currencies "arose as a result of active intervention in foreign exchange markets by the Treasury and the Federal Reserve..." (1983, p. 194). On January 1983, such holdings were \$5.3 billion, primarily composed of German marks, Swiss francs, and Japanese yen. Partee suggested that the substance of the Monetary Control Act and the responsible policy attitude of the Fed were "ample safeguards to prevent section 105(b)(2) from being used by the Federal Reserve as a basis for assisting governments in financial difficulties" (1983, p. 195). Yet in the next paragraph, he admitted that part of the Fed's holdings were Mexican pesos acquired in connection with the Bank of Mexico's drawing on the \$325 million swap

arrangement with the Federal Reserve. The peso holdings were then "invested" in an interest-bearing account at the Bank of Mexico. "When the swap drawing is unwound," Partee assured the committee, "the pesos will be exchanged with the Mexican central bank for dollars at the same rate of exchange at which they were acquired" (1983, p. 195). However, the income report of the Federal Reserve Banks for the year 1983 has the following deduction from current income: "\$456 million of unrealized loss on assets denominated in foreign currencies related to revaluation of the assets at market exchange rates" (*FRB*, 1984, p. 109).

#### IV. Conclusion: The Inevitability of Federal Reserve Power

The overwhelming votes by which the DIDMCA of 1980 passed both Houses of Congress primarily reflected the deregulatory aspects of Title II rather than the augmented powers of the Fed under Title I. Most congressmen simply "checked out" on the vote because they favored more and fairer competition in the financial industry.

Since the Federal Reserve already had complete control over the monetary system, the DIDMCA of 1980 could not augment the Fed's money-creating powers. The Act, however, extended the Fed's regulatory powers over the financial system and it added to the Fed's ability to intervene in foreign affairs by buying and monetizing foreign government obligations. It therefore put the Fed into the role of a silent partner, or even a surrogate, of the State Department for bailing out bankrupt foreign governments who had unmanageable debts due to several large banks in the United States.

Fed officials in their testimony to congressional committees persistently and doggedly advanced one major theme: the Fed had to have more power—to fight inflation, to prevent chaos in the financial industry from deregulation, and to act as an insurance institution for failing banks who might drag other institutions down with them. By misdirection and subterfuge, the Fed inveigled an unwary Congress into doing its bidding. The



Fed's success serves as an object lesson in demonstrating the ineffectiveness of legislative checks on the scope of policy actions assigned to an agency with discretionary powers.

#### REFERENCES

- G. William Miller, "Statement" to Committee on Banking, Housing and Urban Affairs, U.S. Senate, February 26, 1979, *Federal Reserve Bulletin*, March 1979, 65, 229-35.
- J. Charles Partee, "Statement" to Subcommittee on Financial Institutions of the Committee on Banking, Finance and Urban Affairs, U.S. House of Representatives, May 15, 1979, *Federal Reserve Bulletin*, June 1979, 65, 460-67.
- \_\_\_\_\_, "Statement" to Subcommittee on Financial Institutions of the Committee on Banking, Finance and Urban Affairs, U.S. House of Representatives, June 27, 1979, *Federal Reserve Bulletin*, July 1979, 65, 541-44.
- \_\_\_\_\_, "Statement" to Subcommittee on Domestic Monetary Policy of the Committee on Banking, Finance and Urban Affairs, U.S. House of Representatives, March 10, 1983, *Federal Reserve Bulletin*, March 1983, 69, 193-96.
- Paul Volcker, "Statement" to Committee on Banking, Housing and Urban Affairs, U.S. Senate, September 26, 1979, *Federal Reserve Bulletin*, October 1979, 65, 822-29.
- \_\_\_\_\_, "Statement" to Committee on Banking, Housing and Urban Affairs, U.S. Senate, February 4, 1980, *Federal Reserve Bulletin*, February 1980, 66, 143-48.
- Board of Governors of the Federal Reserve System, *Federal Reserve Bulletin*, various issues, 1979-1984.
- U.S. Congress, *Congressional Record*, 125, part 15, 96th Cong., 1st Sess., July 16-July 20, 1979, 18653-19932.
- \_\_\_\_\_, *Congressional Record*, 126, part 6, 96th Cong., 2nd Sess., March 27-April 20, 1980, 6831-8138.



# Deregulation and Monetary Reform

By LELAND B. YEAGER\*

In the past we could distinguish sharply between currency and checkable deposits, on the one hand, and savings accounts and other near- and nonmonies, on the other hand. Unambiguously defining money and measuring its quantity has now been growing more difficult. Inflation-boosted nominal interest rates have promoted wriggling around requirements for non-interest-bearing reserves and around interest rate ceilings. Responses include money market funds, money market deposit accounts, NOW and Super-NOW accounts, overnight repurchase agreements, aspects of the Eurocurrency market, and cash-management accounts offered by brokerage houses. Deregulation has been blurring the distinction between banks and non-banks and between things that do and things that do not function as media of exchange.

The late Robert Weintraub, among others, used to argue (1984, for example) that despite institutional changes, the functional relation between nominal *GNP* and the quantity of money will remain stable—money being properly defined. Adjusted definitions and adjusted demand or velocity functions can continue yielding good fits. Weintraub expected continuing success with money defined as fully checkable accounts in depository institutions plus currency in circulation and nonbank travelers' checks.

But does such a possibility warrant confidence in being able to conduct a quantity-oriented monetary policy from month to month and day to day? What accounts should be deemed *fully* checkable? Which institutions count as *depository* institutions? Figuring out, *ex post*, how money should have been defined and regulated is not the same as knowing how to do so currently. (Thomas Simpson, 1984; John Wenninger, 1984; and Gillian Garcia, 1984, offer recent discussions by Federal Reserve economists.)

Robert Hall (1982) has expressed skepticism about the old idea of quantity control:

...[M]onetary regulations imposed by the American and British governments of the past century create a more-or-less stable relation between a certain class of assets called money and nominal spending..., but different regulations would alter that relation. [p. 1552]

Regulation of financial institutions ...had...implications for the stability of the demand for money. ...[M]ost important, a wide variety of methods of carrying out transactions and holding wealth were regulated out of existence. [p. 1554]

...the money stock itself is a creature of inefficient regulation. [p. 1555]

I do not say that the monetarists were wrong in advocating their steady-growth rule. But their remedy was not taken in due time, the disease has grown more complicated, and their old prescription may no longer be the best.

Rolling deregulation back might conceivably restore stable links between money and other variables. However, deregulation is no mere product of ideology. Regulations have been succumbing to powerful market incentives not easily overcome. Conceivably, also, new monetary linkages might stabilize after the current process of transition has run its course. Who can foretell the future?

## I. The BFH System

Prudence recommends being ready: we should contemplate radical alternatives to our existing and changing monetary regime. Even reforms never adopted may provide contrasts affording insights into our present unsatisfactory method of giving determinacy to our unit of account.

The unit in which we quote prices, express debts and terms of other contracts, and keep accounts is the value of the dollar of fiat money, the scruffy dollar bill. This unit is analogous to units of weight and length and is at least as often used in everyday activities.

\*Auburn University, Auburn, AL 26849.

Yet its size is determined in a haphazard, precarious, and downright preposterous manner.

I wish time permitted detailing the absurdities of making our unit of account be the supply-and-demand-determined value of the unit of the fiat medium of exchange. Supply and demand fail to balance smoothly and continuously because they do not meet on a specific market and determine a specific price. "The money market" is just a figure of speech, not a reality; so monetary disequilibrium gets corrected only in a roundabout and often painful way.

One radical contrast with our existing system appears in what Robert Greenfield and I (1983) have called the BFH system (crediting Fischer Black, Eugene Fama, and Robert Hall for ideas borrowed). The unit of account would no longer coincide with the unit of the medium of exchange. No homogeneous medium of exchange would exist as a possible rival unit. The government would define the new unit, just as it defines units of weights and measures. The definition would run in terms of a bundle of commodities so comprehensive as to have a nearly stable value against goods and services in general. The items in the bundle would be precisely specifiable ones with continuously quoted prices.

The government would exert a nudge against the inertia of old practices by conducting its own transactions and accounting in the new unit. The government is bound to influence a new system by the particular way it disengages from its domination of the current one, so policymakers can hardly avoid considering what sort of reform is desirable. Apart from promoting the new BFH unit, the government would practice *laissez-faire* toward the financial system. It would be forbidden to issue money. (The reform is quite different from the often-proposed composite-commodity money.)

Deregulation would give full scope to innovative financial intermediation. Private enterprises would, in effect, repackage investment portfolios into convenient media of exchange. No one can confidently predict future details. It seems likely, though, that institutions would emerge combining the fea-

tures of today's banks, money market funds, and stock mutual funds. Some holdings in these institutions would presumably be dividend-yielding equity shares; others would be accounts denominated in the BFH unit and bearing interest at competitive rates. In either case, people would not only stipulate prices and payments but also write checks on these holdings in the single, precisely defined, stable unit, *not* in heterogeneous goods and securities.

Would checks drawn on and deposits in financial institutions be redeemable? (The same question applies to currency, for some institutions would presumably issue notes and even coins denominated in the BFH unit.) It is unlikely—because so awkward—that institutions would offer and customers demand redemption in bundles of the actual commodities defining the unit. "Indirect redeemability," as James Buchanan has suggested calling it, is more likely. Competition might lead institutions to cash checks and redeem their currencies and deposits in whatever quantities of gold or of specified securities equalled in total value as many standard commodity bundles as the numbers of units of account denominating the obligations to be redeemed.

It is important that institutions would similarly settle net balances due on account of checks (and banknotes) presented at their clearinghouse. They would transfer gold or securities actually worth as many commodity bundles as the numbers of units of account to be settled. Professionals would make and implement the required calculations every business day, and the ordinary person would no more need to know just what determines the purchasing power of the BFH unit than he needs to know what determines that of the dollar nowadays.

Furthermore, routine settlements at the clearinghouse would provide part of the scope for arbitrage that would maintain the operationality of the unit's commodity definition. Suppose, with Arthur Okun (1981, p. 290), that the standard bundle consists of 1 ball + 1 orange (the principle illustrated is the same with a multi-item bundle). Suppose, further, that market forces make 1 ball worth 3 oranges. Then the BFH unit is worth  $1\frac{1}{3}$

balls and, equivalently, 4 oranges. Prices are *U*.75 for a ball and *U*.25 for an orange, totaling *U*1, as they should. Any discrepancy would provide an opportunity for profitable and corrective arbitrage. Numerical examples could readily be given if space permitted.

Although practices under the BFH system would displace money as we now know it, they would not entail the textbook inconveniences of barter. The advantages of a single definite unit of account and convenient methods of payment would be retained and enhanced. Apart from crucial differences in the unit and in the media for redeeming their obligations, financial institutions would be practicing something similar to free banking under a gold standard. (On the success of free banking in Scotland up to 1845 and for analysis, see Lawrence White, 1984.)

With the unit of account and media of exchange divorced, the unit's value (like the meter's length) would be established by definition, not by regulation of any quantity. Quantities of media of exchange would be limited by the real (unit-of-account) quantities people were willing to hold. Quantities would be constrained in much the way operating for mutual-fund shares nowadays, for nearmoneys until the recent blurring of the distinction between them and circulating media, and for money itself in an individual small country to which the purchasing power of its unit is dictated under an international gold standard. The sizes of the asset sides and the liability and equity sides of institutions' balance sheets would be influenced and reconciled largely by market-determined interest and yield rates received on the institutions' loans and investments and paid to their depositors and shareholders.

Trouble in understanding this demand-side (as well as supply-side) limit to quantities stems, I conjecture, from carrying ideas over from our present system of fiat money, whose unit is also the unit of account. Under this system, expansion of the nominal supply causes the nominal amount demanded to increase also through shrinkage of the unit's real size. Things would be quite different with a unit defined without reference to any medium of exchange.

With quantities of media of exchange determined by demand and supply and with checkable deposits and equity holdings in financial institutions having market-clearing flexible yields or prices of their own expressed in the BFH unit, monetary disequilibrium as we have known it could no longer occur. With the value of the unit of account spared from sometimes coming under strong but sluggishly working upward or downward pressure, painful macroeconomic disorders would be practically forestalled.

Beside macroeconomic advantages, the BFH system would provide the monetary saturation whose absence concerned Milton Friedman in his 1969 article, "The Optimum Quantity of Money." Since media of exchange would bear interest or dividends at competitive rates, high opportunity costs would no longer press holders to spend real resources economizing on cash balances.

A further advantage is absence of any base money distinct from more abundant ordinary money. No longer could scrambles to get out of ordinary money into base money cause panics and deflation. Any distrust would be concentrated on specific financial institutions. Investments in them would decline in price and quantity. Competition would favor the more prudently managed institutions. Deregulation would appropriately extend to abolishing government deposit insurance.

Mention of base money reminds us of perhaps the greatest difficulty with the BFH system, that of making the transition. The appearance of attractive alternatives would collapse the demand to hold money of the present type. Holdings of and liabilities on bank-account dollars would pretty well match. The problem is the collapse of demand for Federal Reserve notes and deposits and Treasury coins. Either this base money would lose its value, expropriating its holders, or else the government would have to replace it by ordinary, interest-bearing, burdensome government debt.

I see no satisfactory answer to this problem of transition. Still, the BFH system is worth understanding for the light it sheds on our existing system. Furthermore, if the existing dollar should be destroyed anyway—perhaps by persisting government fiscal irre-

sponsibility—it would be a shame to reconstruct the same old failed system.

## II. Fisher's Proposal, Modified

The awkwardness of shifting away from government money recommends considering a second-best reform. Abandoning any quantity rule, it would combine a price-index rule with gold redeemability in a manner reminiscent of Irving Fisher's "compensated dollar." (See Fisher, 1920, and compare Willford King, 1948, pp. 209–10. The present proposal differs from Fisher's only in details and in the rationale offered.)

The monetary authority would be required not only to target on a comprehensive price index but also to redeem its money on demand in the (changeable) weight of gold actually worth, at current prices, the bundle of goods and services used in specifying the index. The authority would also be required to issue new money in exchange for the calculated amount of gold, with perhaps a slight spread between its selling and buying prices of gold to cover expenses.

The standard objection to irredeemable fiat money managed so as to stabilize a price index centers on lags and the attendant danger of overshooting and of oscillations around the target. Lags supposedly intervene between an incipient money-supply-and-demand imbalance and its reflection in the price index and between movements of the index and policy responses and their corrective effects. Under the system suggested here, though, two-way convertibility plus arbitrage would keep the dollar always equal in value to the (variable) quantity of gold that was in turn equal in value to a specified bundle of goods and services. The problem of lags would thus be circumvented (or so it seems to me).

Convertibility of the sort described would tie the dollar to the specified bundle. The system would not be a gold standard. Some other commodity, or some one or more securities, might well be the redemption medium instead; gold serves here only as an example. No such self-aggravating and devastating scramble for gold could occur as can occur under an ordinary gold standard with

fractional reserves. Any scramble would reflect itself in an increased price of gold both in money and relative to other goods and services, thereby automatically reducing the physical quantity of gold in which the dollar was redeemable.

Incidentally, the monetary authority would not necessarily be restricted to issuing and retiring money only against quantities of redemption medium offered by or demanded by the public. It could aim open market operations in securities at stabilizing the price index directly and so hold down the volume of actual conversions. Still, two-way convertibility would be available to keep the dollar stable against commodities.

Convertibility of this sort would be more than merely decorative. It would impose an additional discipline on the monetary authority by requiring it actually to *do something* at the initiative of money holders. If excessive money creation had raised the price index from the target level of 100 to 120, people would be redeeming money of \$100 face value in gold (or other redemption medium) quoted at \$120, selling the gold for \$120 and redeeming that money in gold worth \$144, and so on. Threatened with running out of gold, the government would have to buy more. In bidding up the dollar price of gold, it would be tending in that particular way to hold down the physical quantity of gold in which the dollar was redeemable. To avoid further debasing the purchasing power of money and further endangering its gold reserves, however, the government would have to pay for its gold purchases otherwise than by issuing or reissuing money. It would have to raise the necessary funds by cutting expenditures or by taxes or noninflationary borrowing. Such discipline would constrain overissue in the first place.

Monetary management would no longer depend on accurate conceptualization, measurement, and regulation of the quantity of money. The logic of this system of a modified compensated dollar, like the logic of the BFH system, would recommend complete deregulation, including free banking. By the way, the BFH system might be interpreted as a nongovernmental, decentralized, and competitive version of the system just described.

## REFERENCES

- Fisher, Irving, *Stabilizing the Dollar*, New York: Macmillan, 1920.
- Friedman, Milton, "The Optimum Quantity of Money," in his *The Optimum Quantity of Money and Other Essays*, Chicago: Aldine, 1969.
- Garcia, Gillian, "The Right Rabbit: Which Intermediate Target Should the Fed Pursue?," *Economic Perspectives, Federal Reserve Bank of Chicago*, May/June 1984, 15-31.
- Greenfield, Robert L. and Yeager, Leland B., "A Laissez Faire Approach to Monetary Stability," *Journal of Money, Credit and Banking*, August 1983, 15, 302-15.
- Hall, Robert E., "Monetary Trends in the United States and the United Kingdom: A Review from the Perspective of New Developments in Monetary Economics," *Journal of Economic Literature*, December 1982, 20, 1552-56.
- King, Willford I., *The Keys to Prosperity*, New York: Committee for Constitutional Government, 1948.
- Okun, Arthur, *Prices and Quantities*, Washington: The Brookings Institution, 1981.
- Simpson, Thomas D., "Changes in the Financial System: Implications for Monetary Policy," *Brookings Papers on Economic Activity*, 1: 1984, 249-65, followed by comment by Alan S. Blinder and discussion by James Duesenberry, Robert Hall, Benjamin Friedman, Ralph Bryant, and Edmund Phelps.
- Weintraub, Robert, "The New Role for Gold in U.S. Monetary Policy," in Barry N. Siegel, ed., *Money in Crisis*, San Francisco: Pacific Institute for Public Policy Research, 1984.
- Wenninger, John, "Financial Innovation—A Complex Problem Even in a Simple Framework," *Quarterly Review, Federal Reserve Bank of New York*, Summer 1984, 9, 1-8.
- White, Lawrence, H., *Free Banking in Britain: Theory, Experience, and Debate, 1800-1845*, New York: Cambridge University Press, 1984.

# Speculation, Deregulation, and the Interest Rate

By LEONARD A. RAPPING AND LAWRENCE B. PULLEY\*

Interest rates during the 1980's have been at average levels unprecedented in both Europe and the United States since the beginnings of the Industrial Revolution. For those who believe that the Fed controls the interest rate, these high rates are attributed to the October 1979 decision to target bank reserves rather than interest rates. In this view, the Fed succeeded in its effort to indirectly regulate the interest rate by restricting bank reserves, and hence credit. However, because financial innovation (for example, NOW and money market accounts) has greatly weakened the link between reserves and credit, we reject this view. We propose instead that the abnormally high interest rates resulted from portfolio or asset adjustments in response to interest rate deregulation (the October 1979 decision as well as the decontrol of interest rates on deposits) coupled with generalized speculation in asset markets. This explanation will be empirically contrasted with the loanable funds explanation which stresses the flow supply and demand for credit and the Fisher-Friedman explanation which stresses inflationary expectations.

## I. The Theory<sup>1</sup>

When severed from the rate of profit, the rate of interest is sometimes viewed as resulting from a portfolio choice between bonds and money. Under these circumstances, the interest rate will depend among other things on the market's expectation of the future rate of interest. Of course, uncertainty about the future imparts an unavoidable element of arbitrariness in determining the interest rate. However, under normal circumstances the

market will form a conventional view of what the rate of interest will be—based largely on evidence from the recent past. As long as the conventional view prevails, speculators act to stabilize the market. However, when the normal structure is disrupted, destabilizing speculation can drive the prices of assets far outside their conventional range, leading to the breakdown of the convention and to expectations that the unwarranted price movement will continue.

Between 1969 and 1977 the long-term rate of interest on Treasury bonds fluctuated between 6 and 7 percent—due to the Fed's policy of pegging interest rates as much as to the market's belief that such a policy would continue. For eight years the convention held and stabilizing speculators were appropriately rewarded for their activities. But in 1978 the long rate started to drift up, and a year later the Fed renounced the policy of pegging interest rates. Some speculators, whose stabilizing activity had been based on the Fed's policy of targeting interest rates, were driven out of the market entirely. However, because it was understood that the Fed's abandonment of interest rate targets and the continued decontrol of deposit rates were signaling higher interest rates and a recession, many others were converted to destabilizing bears. The long rate both rose and became unstable.

Although this simple Keynesian approach is insightful, classification of assets into the two categories, money and bonds, fails to illuminate the effects of decontrol. For the purpose of analyzing the decontrol of interest rates, the dividing line must be fixed between those assets whose returns were regulated (bills, bonds, and deposits) and those assets in which returns are uncontrolled (gold, foreign exchange, equities, real estate, and real goods).

Rapping (1979) has argued that a form of destabilizing speculation (i.e., seeking short-term profits by betting on self-sustained asset

\*University of Massachusetts, Amherst, MA 01003, and Brandeis University, Waltham, MA 02154, respectively.

<sup>1</sup>The discussion in this section draws selectively on an unpublished paper by Rapping and Stephen Bennett (1983).

price changes and disregarding fundamentals) and generalized inflation are common in periods in which the social conventions governing the distribution of income and wealth erode and distributional conflict results. Under these circumstances, relative prices change and economic growth is slow and problematic. With the combination of conflict, rapidly changing relative prices, and slow growth, speculation abounds. The gambling spirit infects the markets for real estate, equities, collectibles, gold and foreign exchange to name the more prominent speculative vehicles of recent years.

The expectation (if not always the reality) of very high returns now dominates behavior. Under these circumstances, the regulation of interest rates will prevent the self-generating increase in the prices of speculative assets from raising the interest rate. However, once interest rates are deregulated, the return on interest-bearing assets will rise in accordance with the high returns expected on speculative assets. Depending on attitudes toward risk, the interest rate might not equal the anticipated yield on the speculative assets.

We refer to the explanation of high interest rates described above as the portfolio or asset-preference explanation, in which prices move in order to equilibrate the demand for a stock of assets to the available supply. The most widely held alternative view is based on the flow supply and demand of credit or loanable funds. In this approach, the demand for credit depends on prices and the scale of output, while the supply of credit depends on the voluntary savings of the public and on the amount of new credit creation by the banking system.

A third explanation for high interest rates is that the prospect of inflation or expected inflation affects the present by raising the current nominal rate of interest. In this long-run competitive equilibrium view, the interest rate is a result of technology interacting with culture and expected inflation rates, all of which determine the investment and savings decisions of a society. Savings decisions result from the choice between present and future consumption while investment decisions are based on expected profitability. The

expectation of inflation will be impounded in current savings and investment decisions in such a way as to raise the nominal interest rate in proportion to the expectation of inflation.

## II. Methodology and Results

### A. The Model

In expectations form, the basic Fisher equation relating the nominal interest rate,  $R$ , expectations of the real rate,  $r$ , and expectations of the inflation rate,  $\pi$ , can be written

$$(1) \quad R_t = E_m(r_t | \phi_{t-1}) + E_m(\pi_t | \phi_{t-1}),$$

where  $\phi_{t-1}$  represents the relevant information set available at the beginning of period  $t$  and  $E_m$  represents market expectations. Assuming the distribution of  $E_m(r_t | \phi_{t-1})$  to be centered on a constant long-run equilibrium rate,  $\rho$ , with short-term fluctuations given by  $u_t$ , (1) becomes

$$(2) \quad R_t = \rho + E_m(\pi_t | \phi_{t-1}) + u_t.$$

The earliest attempts to empirically examine Fisher interest neutrality have focused on (2) above, regressing nominal interest rates on actual or lagged inflation rates or some measure of expected inflation. The results have not been satisfactory. Coefficients on expected inflation are often significantly less than the predicted value of one, and vary across both time periods and maturities.

In response to mounting evidence that the expected real rate itself is not constant, recent studies have allowed changes in the real rate to be associated with changes in a number of other variables (often in the context of general equilibrium macroeconomic models). John Carlson (1977a) and Vito Tanzi (1980) examine the effects of changes in real activity or business cycle behavior (i.e., shifts in the *IS* curve). In downturns, for example, the (expected) real return to capital falls. If, as a consequence, the demand for loanable funds falls more than the supply, the (expected) real interest rate falls. Of course, if shifts in the *IS* curve can affect real rates, so should

shifts in the *LM* curve. James Wilcox (1983) and others have included money supply growth as a proxy for the well-known liquidity effect. Some have suggested increased money growth volatility as an explanation of high rates, but David Berson (1983) argues that interest rate volatility is the more important volatility measure, particularly given the role of risk premiums in the finance literature.

In this paper we employ the augmented Fisher relationship to examine the hypotheses presented in the introduction and detailed in Section I. The model we adopt is the following:

$$(3) \quad R_t = E_m(r_t|\phi_{t-1}) + \beta_1 E_m(\pi_t|\phi_{t-1}) + u_t,$$

$$(4) \quad E_m(r_t|\phi_{t-1}) = \rho - \beta_2 E_m(\pi_t|\phi_{t-1}) - \beta_3 LIQ_t + \beta_4 ACT_t + \beta_5 \sigma_{R_t} + v_t,$$

where  $E_m(\pi_t|\phi_{t-1})$  is included among the explanatory variables in (4) to allow incomplete adjustment to the level of expected inflation. *LIQ* is a liquidity variable reflecting shifts in the *LM* curve, *ACT* is a real activity variable for shifts in the *IS* curve, and  $\sigma_{R_t}$  is a measure of interest rate volatility to incorporate changes in the risk premium.  $\rho$  is the constant component of the expected real rate. All *beta* coefficients are expected to be positive.  $\beta_1$  is not constrained to unity to allow for tax effects. Combining (3) and (4) yields

$$(5) \quad R_t = \rho + \beta_0 E_m(\pi_t|\phi_{t-1}) - \beta_3 LIQ_t + \beta_4 ACT_t + \beta_5 \sigma_{R_t} + w_t,$$

where  $\beta_0 = \beta_1 - \beta_2$  and  $w_t = u_t + v_t$ . With a suitable proxy for  $E_m(\pi_t|\phi_{t-1})$  (5) can be estimated with ordinary least squares.<sup>2</sup>

<sup>2</sup>Of course, no available measure of  $E_m(\pi_t|\phi_{t-1})$  can be expected to be error free so an errors-in-variables problem exists. However, as Tanzi has shown, the coefficients do not seem to be particularly sensitive to the use of the Livingston survey data and *OLS* or any of several instruments and *2SLS*. Since the Livingston data will be used in this study, we note the finding of Kajal Lahiri and Mark Zaporowski (1984) that the Livingston data consistently underpredict true expectations, in which case coefficients based on the Livingston data are overstated.

We now relate the model described above to the explanations of high interest rates previously discussed. Expression (3) is the simple inflationary expectations model. The liquidity and real activity variables in (4) capture the effects of the Keynesian money-bonds asset choice and changes in the demand for loanable funds. Interest rate volatility is included to control for risk. In Part C below, we present evidence that the inflationary expectations and loanable funds models do not explain interest rates after October 1979. Furthermore, the reported significance of a naive shift variable is consistent with the presence of increased speculative activity.

### B. The Data

The expected inflation series is based on the Livingston surveys computed by the method described in Carlson (1977a, b). The survey represents an average of predictions by economists of the value of the *CPI* six months from the survey months of June and December. Therefore, the nominal interest rates used in this study are the six-month Treasury bill rates prevailing in those months. The period consists of the fifty observations 1959:6 through 1983:12.

The liquidity variable (*LIQ*) is measured as the growth rate in seasonally adjusted and revised *M1* (current dollars) over the most recent quarter. For example, the value corresponding to the June observation is the growth rate implied by the first- and second-quarter observations on *M1*. The measure of real activity used is the growth rate in real *GNP* over the last six months (two quarters).<sup>3</sup> A measure of interest rate volatility,  $\sigma_{R_t}$ , is computed as the standard deviation of monthly observations on three-month Treasury bill rates over the most recent twelve-month period.<sup>4</sup>

<sup>3</sup>In this model, federal tax and spending behavior influence the interest rate by their impact on economic activity. This by no means exhausts the avenues whereby the deficits might affect the interest rate. In particular, deficits could alter interest rate expectations and asset pricing behavior.

<sup>4</sup>The three-month rates were used rather than the six-month rates because the monthly observations on these rates were more readily available. The Livingston



## C. Results

To examine the hypotheses described above, we construct the model in (5) above with a duplicate set of explanatory variables (including the constant) which take on the value zero through 1979:6. With these changes the estimating equation becomes

$$(6) \quad R_t = (\rho + \rho') + (\beta_0 + \beta'_0)E_m(\pi_t|\phi_{t-1}) \\ - (\beta_3 + \beta'_3)LIQ_t + (\beta_4 + \beta'_4)ACT_t \\ + (\beta_5 + \beta'_5)\sigma_{Rt} + w_t,$$

where the primed values represent changes in the coefficients after October 1979. A positive and significant value of  $\hat{\rho}'$  would be consistent with our hypothesis that after October 1979, traditional economic relationships in interest rate determination were replaced by what we have labelled speculative factors. These influences helped keep interest rates abnormally high. (Unfortunately, this test is not very powerful since a positive  $\hat{\rho}'$  could result from some other omitted variable or influence.)

Table 1 presents the results of estimating various versions of (6) above using an iterative maximum-likelihood procedure to correct for first-order serial correlation. Credit controls were in effect during the first half of 1980. Therefore, a dummy variable taking the value one for 1980:6 was included to avoid any distortions introduced by these controls. All variables with the exception of expected inflation were reduced by their sample means. The constant terms can therefore (roughly) be viewed as the real rate when inflationary expectations are zero and  $LIQ$ ,  $ACT$ , and  $\sigma$  take on their sample mean values.

Column 1 of Table 1 gives estimates of the basic Fisher relationship. The estimates in column 2 are from the model in (5) when coefficients are not allowed to adjust in the post-1979 period. The model in column 3

TABLE 1—INTEREST RATE REGRESSIONS SEMIANNUAL: JUNE 1959–DECEMBER 1983

Independent variable	Specification			
	(1)	(2)	(3)	(4)
Constant				
$\hat{\rho}$	2.88 <sup>a</sup> (3.56)	3.33 <sup>a</sup> (3.91)	3.04 <sup>a</sup> (5.73)	2.93 <sup>a</sup> (4.95)
$\hat{\rho}'$	—	—	2.06 <sup>a</sup> (2.64)	5.89 <sup>a</sup> (3.04)
$E_m(\pi_t \phi_{t-1})$				
$\hat{\beta}_0$	.88 <sup>a</sup> (6.00)	.78 <sup>a</sup> (5.34)	.73 <sup>a</sup> (6.17)	.76 <sup>a</sup> (5.84)
$\hat{\beta}'_0$	—	—	—	-.53 <sup>a</sup> (-2.19)
$LIQ$				
$\hat{\beta}_3$	—	-.19 <sup>a</sup> (-4.07)	-.16 <sup>a</sup> (-3.20)	-.13 <sup>a</sup> (-2.22)
$\hat{\beta}'_3$	—	—	—	-.24 <sup>a</sup> (-2.04)
$ACT$				
$\hat{\beta}_4$	—	.026 (.61)	.007 (.16)	.010 (.20)
$\hat{\beta}'_4$	—	—	—	.11 (.92)
$\sigma_R$				
$\hat{\beta}_5$	—	.30 <sup>a</sup> (2.09)	.28 <sup>a</sup> (1.93)	.42 (1.01)
$\hat{\beta}'_5$	—	—	—	.23 (.48)
Dummy (June, 1980)				
$\hat{D}$	-6.35 <sup>a</sup> (-7.75)	-8.52 <sup>a</sup> (-9.28)	-8.26 <sup>a</sup> (-8.36)	-9.91 <sup>a</sup> (-7.17)
$\bar{R}^2$ <sup>b</sup>	.88	.91	.91	.92
$SSE$ <sup>c</sup>	49.2	35.8	33.9	28.5
$D-W$ <sup>d</sup>	2.20	2.16	2.02	1.88
$ARI$ <sup>e</sup>	.76 <sup>a</sup> (8.38)	.81 <sup>a</sup> (9.88)	.61 <sup>a</sup> (5.30)	.62 <sup>a</sup> (5.14)

Note: 50 observations; *t*-statistics are shown in parentheses.

<sup>a</sup>Significantly different from zero at the .05 level (one-tailed test).

<sup>b</sup> $R^2$  adjusted for degrees of freedom.

<sup>c</sup>Sum of squared errors.

<sup>d</sup>Durbin-Watson statistic.

<sup>e</sup>Iterative *ML* estimates of the first-order serial correlation coefficient.

differs from that in column 2 by the inclusion of a dummy variable for the post-1979 period. Column 4 contains estimates for the full model in (6).

Consider the results in Table 1. The unprimed values are consistent with our expectations and with the results of other studies. The one notable exception is the statistically insignificant coefficient on the real activity

survey data are from the Federal Reserve Bank of Philadelphia. The interest rate and money supply data are from the Federal Reserve. The *GNP* figures are from the Department of Commerce.

variable.<sup>5</sup> There is also some evidence of coefficient adjustments after 1979. The adjustment on the *LIQ* variable coefficient ( $\hat{\beta}_3$ ) indicates a possible increase in the importance of the liquidity effect. This lends some support to the conclusion of Richard Clarida and Benjamin Friedman (1984) regarding the contribution of slow money growth to the recent high rates. The negative coefficient adjustment on expected inflation ( $\hat{\beta}_0$ ) is consistent with the view that interest rates have continued to rise (or remain high) in the post-1979 period despite declines in actual and expected inflation. Therefore, it may not be reasonable to posit lagging inflationary expectations as a reason for the high rates.

One potentially troublesome aspect of the results in Table 1 is the large and statistically significant first-order serial correlation coefficient (prior to correction). Although some serial correlation is often present in analyses over time, such large values could indicate the omission of an important variable from the analysis of the post-1979 period. Of course, the dummy variable included for the period is at best an imperfect proxy for the factors we describe in Section I. Furthermore, the residual serial correlation is not a product of the post-1979 period alone. When expression (5) is estimated for the period 1959:6 through 1979:6, the computed serial correlation coefficient is .70.

There is evidence in Table 1 for our assertion that a substantial component of the increase in interest rates since October 1979 cannot be explained by the model in (5), even when all slope coefficients are allowed to adjust. The coefficient  $\hat{\rho}$  is positive and statistically significant. Furthermore, the average of the last nine residuals (observations 1979:12–1983:12) for the model in column 4 is .06. If  $\hat{\rho}$  is constrained to be zero and the model in column 4 is reestimated, the average of the last nine residuals becomes .48. (Based on the residual standard deviations, the standard deviations of both these averages is approximately .25.)

<sup>5</sup>As an additional proxy for real activity, the ratio of employment to non-institutional population was used. The measure is employed by Carlson (1977a), but produced results no different from those in Table 1.

### III. Conclusions

Recent high interest rates are not explained by factors which appear in traditional interest rate models. Similar results have been found in the study of foreign exchange markets where neither the purchasing power parity flow approach nor the interest differential asset approach can account for the behavior of the dollar after 1977. Some have attributed the high value of the dollar to speculative behavior in which decision making is contaminated by a crowd psychology. We contend that while direct speculation in assets fixed in money terms may account for high interest rates, the high rates may also be attributed to generalized speculative activity throughout the economy which pulled interest rates up once they were decontrolled.

Of course, our empirical evidence is only suggestive. With the usual caveats of caution, we conclude that unless speculation is penalized or otherwise deterred, continued deregulation of interest rates could mean continued high interest rates. The excessive speculation of the 1920's was penalized in the bath of red ink caused by the financial market panic of 1929–33. In the 1980's, on the other hand, the Fed has wisely renounced this dangerous medicine for financial overexuberance. Unlike the earlier period, it has actively played its role as a lender-of-last-resort. Of necessity, selective penalties are now needed and the authority to administer them will require some relaxation in the process of financial market deregulation.

### REFERENCES

- Berson, David W., "Money Growth Volatility, Uncertainty, and High Interest Rates," *Economic Review, Federal Reserve Bank of Kansas City*, November 1983, 23–38.
- Carlson, John A., (1977a) "Short-Term Interest Rates as Predictors of Inflation: Comment," *American Economic Review*, June 1977, 67, 469–75.
- , (1977b) "A Study of Price Forecasts," *Annals of Economic and Social Measurement*, Winter 1977, 6, 27–56.

- Clarida, Richard H. and Friedman, Benjamin M.**, "The Behavior of U.S. Short-Term Interest Rates Since October, 1979," *Journal of Finance*, July 1984, 34, 671-82.
- Feldstein, Martin**, "Inflation, Income Taxes, and the Rate of Interest: A Theoretical Analysis," *American Economic Review*, December 1976, 66, 809-20.
- Lahiri, Kajal and Zaporowski, Mark**, "Inflation, Expectations, and the Real Interest Rate," Presented at the meetings of the American Statistical Association, August, 1984.
- Rapping, Leonard A.**, "The Domestic and International Aspects of Structural Inflation," in James H. Gapinski and Charles E. Rockwood, eds., *Essays in Post Keynesian Inflation*, Cambridge: Ballinger, 1979.
- \_\_\_\_\_ and **Bennett, Stephen**, "Reflections on the Interest Rate," Project on Economic Restructuring, University of Massachusetts-Amherst, July 1983.
- Tanzi, Vito**, "Inflationary Expectations, Economic Activity, Taxes, and Interest Rates," *American Economic Review*, March 1980, 70, 12-21.
- Wilcox, James A.**, "Why Real Interest Rates Were so Low in the 1970's," *American Economic Review*, March 1983, 73, 44-53.

## YESTERYEARS' LONG-RANGE PROJECTIONS: A RETROSPECTIVE

### 1985 Projections of the New York Metropolitan Region Study

By DICK NETZER\*

In 1960, the Harvard University Press published *Metropolis 1985*, the final volume presenting and interpreting the results of the New York Metropolitan Region Study, written by the study's director, Raymond Vernon. The study was a three-year effort conducted by Harvard's Graduate School of Public Administration (now the Kennedy School of Government) for the Regional Plan Association of New York. The specific output of the study was twofold: nine volumes of urban analysis and an elaborate set of economic and demographic projections from a mid-1950's base to 1985, to form the foundation for a new regional plan for the 22-county metropolitan region. The first Regional Plan for New York and Its Environs—the first regional plan anywhere—had been produced between 1922 and 1929 and had been founded on a pioneering economic study directed by Robert Murray Haig of Columbia.

The 1956–59 study was one of a series of large-scale regional studies (and much the largest and most ambitious of that series) done in the late 1950's and early 1960's with financing by the Ford Foundation and other foundations and, in reality, inspired by Ford. The methodological contributions of the New York study were considerable: among other things, it pioneered in the regional application of input-output analysis, in regional population projection techniques and in the application of factor analysis to local government expenditure. The study and the contemporaneous work of Edgar S. Dunn, Jr. (see Harvey Perloff et al., 1960) were the first empirical uses of the "shift-share" analysis that became a basic technique of regional economics. Indeed, the study formed a major

component of the technical foundation of the empirical side of urban and regional economics, then in its infancy.

However, the purpose here is not to review the technical contributions of the study. Instead, this paper compares the picture of the New York region painted in the study with the realities as of the early 1980's (usually 1981, to avoid economic data that reflect the deep 1982–83 recession), to address three questions. First, how different are the American economy and the economy of our largest urban area from the forecasts made by the best of analysts a quarter-century ago? Second, was the conception of the process of American urban economic development that lay behind the numbers a valid one? Third, what does this tell us about the utility of long-term projections for public policy?

The cryptic answers to the questions: first, the economic circumstances of the 1980's are quite different from the projected circumstances, almost entirely because of national (not regional) differences, including a significantly lower level of population and aggregate economic activity, and more rapid structural changes—mostly in the directions forecast in 1960. Second, as this implies, the conceptual underpinnings were, with only a few exceptions, spectacularly correct. Third, had public policymakers in the New York area accepted the forecasts as revealed truth and rigidly followed them in policy actions, they would have overbuilt some elements of public capital but otherwise would have done pretty much the right things. Had they accepted the underlying conception but recalibrated the projections periodically to reflect the national aggregates, even the over-capacity errors would have been avoided. This particular set of long-term projections, if no other, could have been useful for public policy, had attention been paid.

\*Urban Research Center, New York University, 4 Washington Square North, New York, NY 10003.

### I. The Nation and the Region

In the nine volumes of the study proper, there is relatively little explicit discussion of the national economy; the formal national projections that were an inescapable starting point for the regional analysis and projections appear in the technical supplement (Barbara Berman et al., 1960), complete with a disclaimer to the effect that the study essentially "borrowed" an existing model of the national economy for the 1955-85 period. That model, like *all* long-term forecasts of that era, projected a great deal of economic growth, to a very considerable extent stemming from the high current birth rates, which yielded high population growth rates. If age- and sex-specific labor force participation rates change only moderately and recent trends in labor productivity are extrapolated, the inevitable result is, at the cyclical peak that occurs at or near the target year, high growth rates in employment and output: in the study, an average annual growth rate of 4.3 percent in real *GNP* and of 2.4 percent in *GNP* per civilian employee over the thirty-year period (Vernon, Table A-1, p. 232).

There is a real problem in filling such large envelopes with output and employment by sector and industry. One cannot simply specify the rise of new, then-nonexistent industries (although one may be sure that some such industries will develop) or forecast the demise of industries that currently show all signs of health. Instead, the responsible course is to forecast that the sectoral composition of the national economy will change slowly, along the lines that then seem apparent. In the mid-1950's, that meant projecting continued growth in the main durable goods manufacturing industries, notably the auto industry, and an approximately stable share of manufacturing in total employment nationally. The results were plausible then, although they seem absurd today. For example, the background tables imply a projected increase in motor vehicle output from 1954 to 1981 (neither of them banner years for the industry) of close to 300 percent in constant dollars; the actual increase was about 100 percent. Most of the actual increase took the form of more value per unit; to realize the

TABLE 1—AGGREGATE PROJECTIONS AND RESULTS  
1980 AND 1981

	U.S.	New York Region
Levels (millions):		
Employment, 1981		
Projected	96.5	9.0
Actual	93.2	7.9
Population, 1980		
Projected	260.6	22.1
Actual	226.5	17.1
Annual Growth Rate <sup>a</sup>		
Employment		
Projected	1.7	1.4
Actual	1.5	0.9
Population		
Projected	1.8	1.5
Actual	1.3	0.5

*Note:* Projections are interpolations between the 1975 and 1985 figures in the appendix tables in Vernon with an adjustment for subsequent revisions in base-year employment data. Actual population figures from the *Census of Population: 1960*. The employment concept is "persons engaged in production," as in the National Income and Product Accounts, the source of the U.S. "actuals." Actual employment for the region based on *County Business Patterns*, supplemented by author's estimates for self-employment (based on the economic censuses) and Bureau of Labor Statistics data on government employment by county.

<sup>a</sup>Shown in percent: U.S. growth rates, 1955-81; New York region employment, 1956-81, and population, 1955-80.

projections, domestic auto output of about 12.5 million units, rather than 5.6 million, would have been needed in 1981.

In any event, the baby boom of the 1950's ended abruptly, and the national aggregates are well below those projected, employment less so than population, which is largely a consequence of there being so many fewer children in the mix today (see Table 1). The New York region had been expected to grow more slowly than the nation—what one would expect of mature regions—but in reality the growth was sharply below the national rates, not 80-odd percent of those rates. But to some extent even the accelerated relative regional declines can be attributed to the nationwide demographic factors. The study forecast that after 1980 the region would experience net out-migration

for the first time ever. Net out-migration appears to have started in the early 1970's; the early start seems partly due to the small numbers of potential *gross* in-migrants to be expected with a smaller national population, as well as the relative attractiveness of the region per se.

However, as Table 1 suggests, there was in fact some relative decline in the region's attractiveness, and more so as a place of residence than as the location of employment. The region's share of the national economy, as measured by employment, did decline somewhat more rapidly than the study had forecast. The region had 10.3 percent of U.S. employment in 1954 and the study (implicitly) projected that the share would fall to 9.1 percent by 1981; the actual result was 8.5 percent. The region's share of nonmanufacturing employment is now very close to the projected share, at 9.0 percent, which is a strong performance given the relative decline in population and the large extent to which local nonmanufacturing employment is local-population-serving. The region's relative weakness has been in manufacturing, in the study projected to decline, in its share of U.S. employment, from 11.6 percent in 1954 to 9.2 percent in 1981, but actually declining to 7.0 percent. As Table 2 shows, for the region as a whole, the actual absolute increase in non-manufacturing employment exceeded the projections, while there was a half-million decline in manufacturing employment instead of a projected 716,000 increase.

As suggested above, the fundamental error in the study's vision of industrial structure was the failure to forecast the sharp change in the role of manufacturing: manufacturing was projected to continue at about 28 percent of total employment from 1954 into the 1980's, but in fact declined to 21.5 percent by 1981. This error affected the region on the "share" as well as the "mix" side. In the study, the strongly growing manufacturing industries that had traditionally been heavily concentrated in the industrial Midwest were expected to grow mainly at dispersed locations, probably on the coasts, with the New York region actually increasing its share of employment in some of these industries, as

TABLE 2—PROJECTED AND ACTUAL EMPLOYMENT  
CHANGES IN NEW YORK REGION, 1956–81  
(IN THOUSANDS)

	Projected	Actual
Entire Region		
Total	2,567	1,531
Manufacturing	716	– 507
Finance and business and professional services	828	1,286
Other industries	1,023	752
New York City		
Total	434	– 315
Manufacturing	49	– 470
Finance and business and professional services	315	449
Other industries	70	– 294
Rest of Region		
Total	2,133	1,846
Manufacturing	667	– 37
Finance and business and professional services	513	837
Other industries	953	1,046

*Note:* See Table 1 for sources of data. Manufacturing excludes "administrative and auxiliary." "Finance and business and professional services" includes finance, insurance and real estate; administrative and auxiliary; radio and television broadcasting; and Standard Industrial Classification codes 73, 80, 81, 82, 84, 86, and 89.

had been the case with regard to the automobile industry in the years immediately preceding the study. In fact, there was much less total growth to be located and, with one major exception that is idiosyncratic to a single firm (Grumman and aerospace), the New York region did not gain from the diffusion of manufacturing away from the Great Lakes.

Manufacturing aside, the industrial structure errors in the projections were ones of degree, not kind: forecasting shifts in the right direction, but understating their magnitude, in particular underestimating the growth in jobs in government, finance, services and communications. The implicit 1981 national projection for services was about one-third below the level realized. Table 2 shows an underestimate of similar proportions for finance and business and professional services in the region. The study volume that focused on the shift-share analysis (Robert Lichtenberg, 1960) concluded that the region had a favorable mix that would be

offset by some competitive loss in share. The reality seems to be that the mix—the heavy concentration on finance and services—was even more favorable than had been forecast, but the loss of share also more pronounced; the job loss associated with the latter factor was compounded by the lesser growth in local population-serving jobs to yield the result shown in Table 2.

## II. Changes within the Region

For much of the study's audience, the aggregate size of the region was not of great moment: the important questions concerned what was likely to be going on within the region, and the study's findings on this score received considerable, often angry, attention. The technicians, however, were well aware that the intraregional projections were beset by serious methodological problems—for example, a miniature version of the shift-share analysis does not really work within regions, but there is no better alternative—and that the likelihood of serious, nonoffsetting errors was great. The test of the accuracy of these projections is not the actual numbers, but whether the vision was valid.

That vision sounds commonplace today, but it was hardly the conventional wisdom in 1960. New York City (except Staten Island, just about to be linked by bridge to the rest of the city) and the other old core cities would decline in population, the inner suburbs would grow moderately and the outer suburbs by very large numbers. The Manhattan central business district (*CBD*) would prosper mightily, and there would be substantial job growth everywhere outside the other parts of the core. The real difficulties would be in the vast expanses between the *CBD* (and the already-existing close-in slums) and the post-World War II suburban development, the sections Vernon called the "gray areas." In the late 1950's, these areas were occupied by manufacturing and other goods-handling activities that appeared to be viable and by what seemed stable middle- and working-class residential neighborhoods, to a considerable extent housing constructed during the great boom associated with the building of the subway network to Brooklyn, the

Bronx and Queens, from about 1910 through the 1930's. As Vernon pointed out, by 1985 that housing would be 50-or-more years old; it was built to not very high standards to begin with, at often very high densities (especially in the Bronx), and was likely to be seen by ordinary people in the 1980's as distinctly inferior to the characteristic suburban housing and neighborhood patterns developed in more recent decades. The neighborhoods might be an improvement for recent slum dwellers, but they would be everybody else's third choice, with vacancies and undermaintenance the result. The industrial areas also would be seen as inferior choices, inferior to newer industrial areas on the fringes of New York City (for example, near what is now Kennedy airport), to the outlying parts of the region and even to the industrial pockets within the *CBD*; the worst-faring industrial areas would be those in the smaller core cities of the region—Newark, Jersey City, Paterson, Bridgeport, and the like.

Those forecasts, denounced at the time, alas have proved to be, on the whole, optimistic. The smaller industrial cities have had massive losses in manufacturing employment and little if any offsetting increases in other types of employment. Within New York City, there was (between 1959 and 1981) a decline in manufacturing employment of roughly equal proportions—about 50 percent—in the *CBD* and in the outer boroughs. In the *CBD*, that job loss was made up by increases in service-exporting activities, with stability in total employment in local-population-serving activities. But in the Bronx, Brooklyn, and Queens, jobs in the local service activities declined and jobs in service-exporting activities are negligible, aside from the 50,000 or so in air transportation in Queens, with the upshot heavy losses in total employment (about 200,000 in Brooklyn alone), far heavier than the study projected. (For 1959 employment data, see Regional Plan Association, 1967.)

Similarly, the projections for population (and housing) in the gray areas were not dire enough. The study forecast that population in the three "old" boroughs of New York City (Manhattan, the Bronx, and Brooklyn) would decline considerably, but the decline

TABLE 3—PROJECTED AND ACTUAL POPULATION  
AND EMPLOYMENT CHANGES BY SECTOR OF REGION  
(IN THOUSANDS)

	Projected	Actual
Population Change, 1955–80		
Region	7,013	2,015
New York City	–84	–735
East Counties <sup>a</sup>	1,178	1,116
North Counties <sup>b</sup>	2,048	748
West Counties <sup>c</sup>	3,871	886
Employment Change, 1956–81		
Region	2,567	1,531
New York City	434	–315
East Counties <sup>a</sup>	346	595
North Counties <sup>b</sup>	572	494
West Counties <sup>c</sup>	1,315	756

Note: See Table 1 for sources of data.

<sup>a</sup>Nassau and Suffolk counties.

<sup>b</sup>Five counties in New York State and Fairfield County, Connecticut.

<sup>c</sup>Nine counties in northeast New Jersey.

was about 500,000 greater than forecast. The decline in household size was not sufficient to sustain housing demand outside of Manhattan, where the number of occupied housing units in 1980 was roughly equal to the 1960 level. However, in the Bronx and Brooklyn, there was a substantial decline—about 57,000—in housing units: housing investment in a few Manhattan-like pockets (Riverdale in the Bronx and the brownstone areas of Brooklyn) was far from enough to offset the massive abandonment in the South Bronx and eastern Brooklyn, in the former case including abandonment of what had been up-market middle-class housing built in the 1920's and 1930's.

Projected and actual changes in population and employment in the region outside New York City are shown in Table 3. The region's population and employment are somewhat more decentralized than had been projected, but the more interesting story seems to be the differences among the sectors, defined here roughly by compass directions. Long Island (in East Counties) actually attracted substantially more jobs and roughly the same number of residents as had been projected, but the other sectors grew much more slowly than projected. In part, the simple absence of population pressure

explains the results: with fewer people seeking places to live, one would expect less overspill from inner to outer areas. Moreover, restrictive land use controls should hold up better in the absence of such market-generated demand pressure: the study's projections implied that low-density zoning in outer areas, like northern Westchester and Connecticut, would not hold up at all.

The disparity between projected and actual is especially wide for the New Jersey sector (in West Counties), including both inner and outer sections. This seems related to the slow growth of manufacturing employment in general. The study's picture of the region shows New Jersey as having very large numbers of households supported by well-paid locally based blue-collar jobs in manufacturing. The counties to the north, especially Fairfield County in Connecticut, also attracted fewer manufacturing jobs than expected, but were more successful than New Jersey in attracting service-exporting activities.

### III. Projections, Perceptions, and Public Policy

On balance, therefore, it seems clear that the picture of the region in the early 1980's painted in the study was accurate in many respects, and would have been very accurate indeed had a smaller national population and the rapid shift away from manufacturing nationally been forecast. *Metropolis 1985* ends with a passage predicting that most of the region's residents would consider themselves reasonably well off in 1985, certainly as compared to 1960, while the planners and other experts would be highly dissatisfied. Surely, that is not the case: the "subjective reality," the perceptions of ordinary people, may be a lot worse, not much better, than the experts' perceptions of "objective reality." The old cities are worse off than expected by a wide margin; the public capital stock is in bad shape relative to the implicit forecasts in the study; and real income levels are a good deal lower than projected, to a considerable extent because of the macroeconomic events of the 1970's.

The New York study was not alone in seeming overly optimistic, after the event:



most long-term urban and regional studies from the late 1950's into the early 1970's, in all parts of the United States, share that characteristic, which seems to be largely responsible for the current shunning of long-term projections by urban analysts. However, the fact that the study was overly optimistic does not explain why its projections were largely ignored by policymakers in New York. Instead, the study's findings were rejected as too *pessimistic*. And so the city government spent a good deal of money pointlessly, in rehabilitating high-density housing in depopulating sections (rather than encouraging density-reducing land-use changes), in expanding subway service between Brooklyn and Manhattan, and in developing still-unused industrial parks to accommodate all that nonexistent manufacturing growth, to cite some examples. Moreover, the study implied that, in the core and the older suburbs, there would be need for an explicit strategy to replace infrastructure over time, quantified in a sequel to the study (Regional Plan Association, 1962). Historically, American state and local governments never explicitly replaced obsolete facilities, but instead did so as a by-product of building the new facilities necessary to accommodate population growth. Without such growth, replacement by inadvertence does not work, and did not work in the event.

No doubt, any long-term urban projections done today in the United States would be guarded indeed. But if the 1956-60 New

York study is any indication, intelligent, internally consistent projections that provide a coherent vision of the future of an urban region, whether that vision is upbeat or downbeat, can be of real use in public policy formation. They also can contribute to the development of analytic tools; no urban economists will allege that we have all that we need today.

## REFERENCES

- Berman, Barbara R., Chinitz, Benjamin and Hoover, Edgar M., *Projection of a Metropolis*, Cambridge: Harvard University Press, 1960.
- Lichtenberg, Robert M., *One Tenth of a Nation*, Cambridge: Harvard University Press, 1960.
- Perloff et al., Harvey S., *Regions, Resources, and Economic Growth*, Baltimore: Johns Hopkins University Press, 1960.
- Vernon, Raymond, *Metropolis 1985*, Cambridge: Harvard University Press, 1960.
- Regional Plan Association, *Spread City*, New York: The Association, 1962.
- \_\_\_\_\_, *The Region's Growth*, New York: The Association, 1967.
- U.S. Department of Commerce, Bureau of the Census, *County Business Patterns, 1981: United States, Connecticut, New Jersey and New York*, Washington: USGPO, 1983.
- \_\_\_\_\_, Bureau of Economic Analysis, *Survey of Current Business*, July 1984.

# Long-Term Forecasts in International Economics

By WILLIAM R. CLINE\*

Eric Blair's forecast from the 1940's called for a 1984 international political economy with constant war between three global powers, designed to use up resources otherwise so abundant that the masses would be free to reflect on, and consequently depose, the bureaucratic elite (George Orwell, 1949). Orwell's forecast was wrong, although perhaps in part only because it was a self-negating prophecy: one that induces corrective measures.

Undoubtedly the most successful long-term economic forecast was Joseph's prediction to the Pharaoh of the fourteen-year agrarian business cycle. Others with less privileged information have fared less well. A founder of our own profession, Malthus, erroneously predicted long-term stagnation at the subsistence level, because he underestimated technical change (although for Sub-Saharan Africa and parts of South Asia, the Malthusian projection appears closer to the mark). This essay reviews long-term forecasts from the last several decades in the area of international economics, to see whether patterns can be distinguished in their success or failure.

## I. Secular Terms of Trade

In the 1940's and 1950's, some prominent economists maintained the thesis of a secular decline in the terms of trade for raw materials. Ragnar Nurkse (1944) spoke of "chronic" depression in terms of trade for agricultural goods, and Raul Prebisch (1959) built a theoretical justification for import-substituting industrialization around the concept that unfavorable demand elasticities would condemn producers of raw materials to continu-

ously declining terms of trade if they did not diversify toward industry.

The record has shown that terms of trade have fluctuated but not declined secularly. Declining terms of trade may be shown by selecting a high base period (such as the Korean War years) and a low terminal period (such as 1982); rising terms of trade may be shown by choosing a low base (the 1930's) and a high terminal period (1973-74). Technical change has lowered real prices of industrial goods, tending to offset lower income elasticities for raw materials. And there may be demand asymmetries that cause much sharper price increases for raw materials during boom years than the more gradual declines during years of weak international growth.

It is difficult to judge whether the long-term forecast of declining terms of trade has on balance done policy damage. It probably induced excessive and inefficient industrialization in developing countries. But some countries have succeeded in using the resulting industrial base for subsequent exports on an internationally competitive basis. Moreover, there is at least some degree of self-negation in the forecast: the fact that so many developing country planners accepted it meant less investment in raw materials exports, and thus a more favorable trend in terms of trade, than otherwise would have occurred.

## II. The Dollar Shortage

The record is less ambiguous on another popular long-term forecast of the early post-war period: that the dollar would be in chronic shortage. Defining the dollar shortage as a persistent tendency for the U.S. current account surplus to exceed its long-term capital outflow, Charles Kindleberger (1950) maintained that the shortage had existed since 1914 (except for the period of normalcy in 1922-28). He attributed the

\*Institute for International Economics, 11 Dupont Circle, NW, Washington, D.C., 20036.

problem primarily to a tendency toward secular stagnation in the United States relative to the rest of the world (p. 170).

Writing in 1956, Donald MacDougall (1957) acknowledged that there had been a "striking improvement" in the dollar shortage problem since 1946 as the result of increased production and currency devaluations in Europe. Yet he warned that dollar shortage was likely to return because of periodic U.S. recession, the tendency of other countries to live beyond their means as they sought to emulate U.S. living standards, and the structural advantages in U.S. trade in manufactures. His central forecast was that the rest of the world would run annual deficits with the United States, rising from \$ $\frac{1}{2}$  billion to \$4 billion over a decade. Because this trend would exhaust reserves in the rest of the world, MacDougall predicted a currency crisis in "five or six years at the latest" (pp. 325-26).

Even as authors were devising structural explanations for the dollar shortage, however, it was disappearing. After a large surplus on basic balance of payments (current account plus long-term capital and aid) of \$3.8 billion in 1947 (24 percent of merchandise exports), the balance shifted to a deficit of \$3.4 billion (34 percent of exports) during the Korean War year of 1950, the very year Kindleberger's book was published. Although the deficit then moderated, over the six-year period of MacDougall's predicted dollar shortage crisis (1957-62), the U.S. basic balance was in deficit by an average of \$2.3 billion (12 percent of exports; Walter Salant et al., 1963, pp. 6, 278-80). By the turn of the decade Robert Triffin (1960) was pointing out that the ongoing U.S. deficit would undermine the gold-exchange system by removing credibility of U.S. conversion of dollars into gold.

It is unlikely that predictions of chronic dollar shortage caused much policy mischief, however. The policy implications—increased U.S. economic assistance abroad, intensified efforts at export development in Europe, and so forth—can only be judged to have been salutary. But the predictions do illustrate one of the principal risks of long-term forecasts:

the tendency to project the conditions of the past into the distant future.

### III. Raw Materials Shortages

Richard Cooper (1975) has reviewed another long-term forecast: the projections in 1952 of the President's Materials Policy Commission (Paley Commission) on supply of raw materials to the U.S. economy, a report prompted by concern during Korean War scarcities. Cooper found that the Paley Commission projected excessive requirements of raw materials, with the mean forecast (among 24 mineral products) 46 percent above actual U.S. consumption in 1972. Moreover, the Paley Commission underpredicted U.S. growth at 2.8 percent annually instead of the actual 3.6 percent, for a total GNP shortfall of 20 percent by 1975, with underprediction evenly divided between population and productivity growth. Accordingly, its projections of raw materials use relative to GNP were even further off the mark. After finding that movements in relative prices could not explain the shortfall of minerals consumption from the forecast, Cooper concluded that the Paley Commission seriously underestimated technological change.

In one qualitative dimension, the Paley Commission was more successful: it correctly predicted that "It will pay the United States to import much more raw material in the future than in the past" (Vol. I, p. 12). Its projections expected the 1952 Korean War surge in import shares for copper, lead, and zinc to become permanent by 1975. But the authors missed the most important case of rising import dependence; they erroneously expected the nation still to be almost self-sufficient in oil by 1975, despite their unusually accurate forecast of total energy needs (Paley Commission, Vol. I, p. 22; Cooper, p. 241).

The Paley Commission projections illustrate two ways that projections can go wrong: inappropriate parameters in the projection model, and erroneous assumptions about the exogenous variables. (A third source of error would be inaccurate model

structure, as opposed to parameter values.) The Paley Commission's mistake on the exogenous variable, economic growth, was understandable following one decade of depression and another in which a common view was that peacetime demand would be slack. Less obvious, but actually more important for the error in the outcome, was parameter error. From Cooper's summary data, it may be inferred that the mean elasticity of minerals consumption with respect to *GNP* (among 24 products) was assumed to be 1.1, and the average 1.2 (excluding one outlier, manganese). In actual experience the elasticities turned out to be 0.74 and 0.85, respectively.

This divergence illustrates another aspect of long-term forecasting. It should be possible to inform the elements of the forecast by theory. In retrospect, theory could have suggested to the Paley Commission experts that technical change (as well as a secular rise in the share of services) would mean a less than unitary elasticity of raw materials inputs with respect to *GDP*. Although the Paley Commission stated that it had considered both factors (Vol. II, p. 112), its implicit elasticities indicate that insufficient weight was given to technological change in particular.

From a policy perspective, it is tempting to conclude that the projections were formulated to be on the safe side, in an environment of concern about security of access to raw materials. There is something to be said in favor of this direction of bias, especially if it leads to heightened awareness but not to costly and unnecessary remedial efforts.

#### IV. Balance of Payments

In 1963, the Brookings Institution published a projection of the U.S. balance of payments in 1968 (Salant et al.). The U.S. payments position and the dollar had come under pressure beginning in 1958, and the weak dollar was seen as a constraint on fiscal-monetary policies for full employment. An important policy question was whether the situation would be self-correcting. The projection effort was heroic, because the balance of payments concerns differences between trade (and financial) credits and debits, and even modest errors in the underly-

ing variables can yield large errors in the differences.

The Brookings team constructed a three-country model (United States, Europe, and rest of the world). The U.S. exports depended on Europe's income, its prices relative to U.S. export prices, and on rest of world export earnings. The U.S. imports depended on domestic income and prices relative to European export prices. The authors assumed slightly higher U.S. growth than for Europe (4.8 percent vs. 4.2 percent). But their critical assumption was that U.S. export prices would rise at only 0.5 percent, while Europe's would rise at 1.5 percent. Their rationale was that declining labor force growth in Europe would put upward pressure on wages (whether the guest workers were taken into account is unclear), and that European governments would not tolerate much unemployment to avoid inflation. Primarily because of improving U.S. price competitiveness, the authors concluded that the U.S. basic balance would shift from a deficit of \$0.85 billion in 1961 to a surplus of \$1.9 billion in 1968.

At least some commentators at the time suggested that by taking price increases as given for the United States but deriving higher price increases from expected wage increases in Europe, the authors were biasing the forecast toward optimism (John Williamson in Joint Economic Committee, 1963, pp. 851-59). But by far the major upset to the forecast was the Vietnam War, which caused a surge in U.S. demand and inflation. Actual export price increases averaged 2.0 percent for the United States and only 0.5 percent for Europe over 1963-68, reversing the projected relationship; and nominal U.S. *GNP* rose 46 percent instead of the projected 36 percent (International Monetary Fund, 1978). The actual outcome was a basic balance deficit of \$6.6 billion (Commerce Department, 1970, pp. 36-67), driven by large private and official long-term capital outflows. The projection's large surplus on goods and services (\$9.1 billion) failed to materialize (the actual figure was \$1.8 billion), leaving a large deficit instead of the projected basic balance surplus of \$1.9 billion (Salant et al., p. 216).

In an illustration of the difference between model error and error in assumptions on exogenous variables, Salant et al. acknowledged "...the possibility that the basic assumptions may be wrong—which is indeed a hazard, but one of forecasting, not of projecting..." (p. 227). And the underlying Brookings model appears not to have been seriously biased; for example, the model captured relatively greater response of imports to income in the United States than in Europe (p. 90), a stylized fact from later econometric work (Hendrik Houthakker and Stephen Magee, 1969). Nonetheless, it is perhaps too generous to grant total absolution for "forecasting" as opposed to "projection" (model as opposed to assumption) error, especially if the intended audience includes policymakers. It would have been more prudent to include as an alternative variant one in which U.S. inflation was significantly higher than that in Europe. Instead, the alternative considered permitted only modest changes (slightly lower U.S. growth, slower growth and lower inflation in Europe—but not faster U.S. export inflation than in the base case), yielding a small basic balance deficit for 1968 (\$0.6 billion).

In retrospect one also wonders whether what might be called the Norman Vincent Peale syndrome was involved: the temptation toward positive thinking in policy analysis. Rephrased as the Burt Lance principle ("if it ain't broke, don't fix it"), this approach is not necessarily wrong, and it may avoid costly measures that turn out to be unnecessary. But emphasizing the positive should be recognized as the opposite of a minimax strategy that errs in the direction of caution to avoid large damage. And to the extent that the Kennedy tax cut and, more seriously, the Johnson spending surge, were facilitated by upbeat forecasts on the balance of payments, the costs of error were potentially significant. In their defense, of course, the authors cannot be blamed for failing to predict the Vietnam War.

## V. Oil

In 1970, then-Secretary of Labor George Schultz headed a government study on the

issue of eliminating oil import quotas (Cabinet Task Force, 1970). The group favored import liberalization for economic reasons and argued that any security risks from greater reliance on imported oil were modest. The authors projected a decline in U.S. oil prices from the protected level of \$3.30 per barrel to the world price of \$2.00 (1969 dollars). They expected these real prices to hold through 1980, or even weaken as new production came in from the Alaskan North Slope. They projected that liberalization would increase the share of imports in 1980 U.S. oil consumption from one-fourth to one-half. But they anticipated little security difficulty because fully half of imports could come from Canada and Latin America, and only one-third would come from more unstable Arab states (with one-eighth from "reliable" Iran).

The task force did not err greatly in its volume estimate for U.S. oil consumption in 1980 (at 19.3 million barrels per day compared with 17.1 mbpd actual). Projected imports were somewhat further above actual (9.8 mbpd vs. 6.9), as some adjustment to sharply higher prices occurred. However, the Western Hemisphere turned out to supply only about one-fifth of imports (with Canadian supply far below projection), while Arab states provided nearly one-half (Commerce Department, 1983, pp. 528, 724–25).

Along with most other observers, the Cabinet Task Force failed to predict the OPEC price revolution. The actual 1980 price of \$34 per barrel was 8 times the projected level in real terms. The uncanny feature of this projection failure is that the task force was aware of the risk from OPEC price boosts, but rejected the risk as minor. The Cabinet Task Force study acknowledged:

Producing countries have sought and will doubtless continue to seek to increase their oil revenue by raising their taxes and royalties.... Whether they succeed in creating an effective price-raising cartel will, however, be largely independent of our import policy. ...[T]he world market seems likely to be more competitive in the future than in the past because the growing number and diversity of producing coun-

tries and companies make it even more difficult to organize and enforce a cartel. [p. 21]

Here the Norman Vincent Peale syndrome was, in retrospect, rampant. Moreover, the analytical framework of the authors based expected price on marginal cost of production and distribution, which they expected to be declining. Yet oil today is priced by the largest suppliers not at marginal production cost but at a reservation price, just as holders of gold bullion do not sell it for the marginal cost of removal from their vaults. Moreover, the task force seems to have been unaware that even as it was writing a historic change was taking place: U.S. proven reserves began to decline substantially (Commerce Department, 1983, p. 730). And the authors made the same mistake repeated frequently by others later in underestimating the cost of "backstop technology," anticipating that synthetic oil from shale could be produced at \$3.00 per barrel (p. 55).

Perhaps most seriously, the authors did not consider the interaction between their own policy recommendation and the likelihood of OPEC cartel action. Their projections indicated that, with liberalization, U.S. oil imports would surge from 2.5 mbpd in 1968 to 9.8 mbpd in 1980 (reaching 5 mbpd even without liberalization). Yet they treated the United States as a marginal importer and the world supply of oil as infinitely elastic. In fact, U.S. imports rose briskly to a peak of 8.5 mbpd in 1979. This increment amounted to perhaps one-fourth of OPEC export capacity. Rising U.S. imports almost certainly strengthened the demand conditions for the sharp OPEC price increases of 1973-74 and 1979-80. Moreover, although the authors were aware of another underpinning of potential OPEC power, low price elasticity of demand (the report used an elasticity of 0.1), they failed to draw the conclusion of considerable risk of a sharp price runup through coordinated OPEC action.

The Cabinet Task Force advocated eliminating oil import quotas because liberalization would save consumers \$5 billion annually (and provide net welfare gains of \$1.5 to \$2 billion annually). In the event, political

power of U.S. oil producers delayed liberalization until the mid-1970's. Actual costs to U.S. consumers from the increase in world oil prices by 1980 was approximately \$90 billion annually at 1969 prices.<sup>1</sup> Thus, if the probability was 5.6 percent (\$5 billion/\$90 billion) or higher that the elimination of U.S. oil import protection would ensure viability of the OPEC price revolution (by strengthening demand for OPEC exports), the recommendation to liberalize was a mistake. The odds are that this probability was at least this high, and in retrospect it would have been better to continue oil protection, although switching from a quota to a tariff would have transferred rents from foreign producers to the U.S. government.

## VI. Other Examples

In the early 1970's steel producers around the world adopted major expansion programs, encouraged by temporary shortages in the 1973 boom. Within a decade there was massive excess capacity in world steel and the principal policy question was how to phase down production capacity. In the late 1970's, developing countries borrowed, and foreign banks lent, heavily on the basis of the implicit long-term forecast that the real interest rate would remain low or negative and export growth would remain rapid. By 1982 the U.S. macro policy mix and global recession thwarted both predictions and the debt crisis resulted.

In 1978, the World Bank issued its first *World Development Report* (World Bank 1978). The authors predicted that, after disappointing growth of 2.8 percent annually in 1970-75, the industrial countries would grow at 4.2 percent in 1975-85. The developing countries, which had managed to continue high growth of 5.9 percent in 1970-75 (in part by borrowing in response to the oil shock), would continue their growth at 5.7 percent in 1975-85 (p. 27). The actual record of growth in 1975-84 turned out to be 2.9

<sup>1</sup>Applying the GDP Deflator to \$34 per barrel, subtracting \$3.30 per barrel as the comparison price, and multiplying by 19 mbpd projected consumption.

percent annually for industrial countries and 3.2 percent for developing countries (IMF, 1984, pp. 169–70). The report did not anticipate the second oil shock, the concerted contractionary policy response to it, and the worst global recession since the 1930s.

### VII. Lessons

A broad implication of this review is that long-term forecasts are rarely right. One might have expected that by random chance at least one-half of the forecasts would turn out to be correct in at least qualitative terms. Yet the sample here is not (intentionally) biased toward historical mistakes.<sup>2</sup> It is possible that an asymmetry exists between successful and unsuccessful long-term forecasts. Both types come in two varieties: adverse and favorable. A successful (accurate) forecast that is adverse may become self-negating as public and private actors take corrective measures; one that is favorable may generate market responses (to profit opportunities) that also reduce its *ex post* validity. There are no corresponding mechanisms for turning bad forecasts into good ones, except random chance. A simpler interpretation, of course, is that long-term forecasts are usually wrong because they are difficult to make. Despite their difficulty, they cannot be avoided; anyone planning for retirement at least implicitly makes them, corporations need them, and governments must make choices that require them.

Long-term forecasters should avoid the tendency merely to extrapolate past experience, a pattern present in nearly all of the forecasts considered here. Attention to likely structural changes in the future, especially technological change, can help avoid this pitfall. In modeling terms, care must be used

not only in model structure, but also in changes from past trends with regard to assumptions on future values of exogenous variables and choice of parameter values. Where possible, theory should be coupled with empirical experience for improved results in all three areas.

Forecasters can take some solace from the likelihood that, in a good forecast, the basic outcome may be predicted correctly even if not all of the underlying variables turn out as expected. If errors are randomly distributed, the chances are that the value of the dependent variable in a multivariate model will not be far wrong, because departures in one direction for some exogenous variables will tend to be offset by departures in the other direction for others (see my 1984 study, p. 169). Finally, policy forecasting should avoid at least two temptations: to corroborate the view of the established regime, and to err systematically on the side of optimism. Ideally, sensitivity analysis should be included, with an assessment of the probabilities of alternative outcomes, and attention to the policy costs of premising action on the wrong one.

### REFERENCES

- Cline, William R., *International Debt: Systemic Risk and Policy Response*, Washington: Institute for International Economics, 1984.
- Cooper, Richard N., "Resource Needs Revisited," *Brookings Papers on Economic Activity*, 1:1975, 238–45.
- Houthakker, Hendrik S. and Magee, Stephen P., "Income and Price Elasticities in World Trade," *Review of Economics and Statistics*, May 1969, 51, 111–25.
- Kindleberger, Charles P., *The Dollar Shortage*, New York: Wiley & Sons, 1950.
- MacDougall, Donald, *The World Dollar Problem*, London: MacMillan, 1957.
- Nurkse, Ragnar, *International Currency Experience*, Princeton: Princeton University Press, 1944.
- Orwell, George, 1984, New York: Harcourt, Brace, Jovanich, 1949.
- Prebisch, Raul, "Commercial Policy in the Underdeveloped Countries," *American Eco-*

<sup>2</sup>Surprisingly, one of the more heavily criticized projections—secular decline in terms of trade—turns out to have been among the more accurate, if allowance is made for the influence of self-negation. Note that a proper test of the hypothesis would involve a simulation of what the terms of trade would have been in the absence of structural transformation away from raw materials exports by policy intervention in developing countries.

- conomic Review Proceedings*, May 1959, 49, 251-73.
- Salant et al., Walter, *The United States Balance of Payments in 1968*, Washington: The Brookings Institution, 1963.
- Triffin, Robert, *Gold and the Dollar Crisis*, New Haven: Yale University Press, 1960.
- Cabinet Task Force on Oil Import Control, *The Oil Import Question: A Report on the Relationship of Oil Imports to National Security*, Washington: USGPO, 1970.
- International Monetary Fund, *International Financial Statistics*, May 1978, 31.
- Joint Economic Committee, *The United States Balance of Payments*, Washington: US-GPO, 1963.
- The President's Materials Policy Commission (Paley Commission), *Resources for Freedom: Vol. I, Foundations for Growth and Security; Vol. II, The Outlook for Key Commodities*, Washington: USGPO, June 1952.
- U.S. Department of Commerce, *Statistical Abstract of the United States 1982-83*, Washington, 1983.
- \_\_\_\_\_, *Survey of Current Business*, March 1970, 50.
- World Bank, *World Development Report, 1978*, Washington: World Bank, 1978.



# The Economic Thought of George Orwell

By JENNIFER ROBACK\*

Despite the volumes that have been written about George Orwell during this past year of 1984, virtually no one has commented on the economic assumptions implicit in his work. Yet understanding the economics in Orwell's vision of hell is necessary if we are to fully appreciate the horror of that vision. Orwell believed that socialism led to totalitarianism; indeed, that is the chief message of Orwell's 1949 classic. But he also believed that capitalism led to breadlines and poverty. This combination of beliefs led him to that profoundly disturbing view he described so vividly in *1984*.

But probably the most remarkable aspect of Orwell's economic thought is that he said so little explicitly on the subject. He wrote literally volumes of social and political commentary, with an economic worldview implicit in every line. Yet, explicit discussion of economics is limited to a few pages in *1984* and offhand comments scattered throughout his book reviews and essays. The statements that he did make suggest that his economic ideas were obviously influenced by Marxist theory, but were not revolutionary. That is, his views were quite conventional for his time.

His views about technology are quite another matter. Orwell was not opposed to or in favor of technology per se. His primary concern was over who would control technology. Would new technologies be widely available for personal use or would they be centrally controlled? In this area Orwell was a decentralist, radical for his time and for ours.

## I. Orwell's Economics

The first thing to be said about Orwell the economist is that he was a socialist. He be-

lieved that income inequality was an inherent feature of capitalism, and this was his primary motivation for becoming a socialist. "I became pro-Socialist more out of disgust with the way the poorer section of the industrial workers were oppressed and neglected than out of any theoretical admiration for a planned society."<sup>1</sup>

Orwell believed that overproduction was a necessary feature of industrial capitalism. This creates a surplus of goods which workers would not be able to afford and which capitalists could not consume. This in turn requires either some means of disposing of the surplus production or a periodic crash of the economic system.

Ever since the end of the nineteenth century, the problem of what to do with the surplus of consumption goods has been latent in industrial society. ...[D]uring the final phase of capitalism roughly between 1920 and 1940[,] [t]he economy of many countries was allowed to stagnate, land went out of cultivation, capital equipment was not added to, great blocks of the population were prevented from working and kept half alive by State charity...the problem was how to keep the wheels of industry turning without increasing the real wealth of the world. Goods must be produced, but they need not be distributed.<sup>2</sup>

Orwell also believed that capitalism had strong tendencies toward monopoly and increased concentration. This exacerbates the overproduction problem. Capital becomes

<sup>1</sup>See, Sonia Orwell and Ian Argus, *The Collected Essays*, ..., Vol. III, p. 403. All quotations in the text are from the *Essays* unless otherwise noted.

<sup>2</sup>Orwell (1983, pp. 155-57). This passage comes from the "forbidden book" of Emmanuel Goldstein. In the context of the story, *The Book* is clearly speaking for Orwell when it describes the capitalist system that formerly existed in Oceania.

\*Assistant Professor of Economics, Yale University, New Haven, CT 06520.

concentrated in fewer and fewer hands, hence the output is owned by a smaller number of people, who are less likely to be able to consume all of it: "The notion that industrialism must end in monopoly, and that monopoly must imply tyranny, is not a startling one" (*Essays*, Vol. IV, p. 163).

These ideas, overproduction and periodic crashes, the tendency toward monopoly, and income inequality are traditional socialist concerns. And on the basis of these beliefs and concerns, Orwell was a socialist. This general worldview was widely held by British intellectuals of Orwell's time. Indeed, the period after World War I was one of the highwater marks for socialism, both in Britain and in the United States. It is in this sense that Orwell's economic views can be described as quite conventional for his time.

However, Orwell was no ordinary socialist, once we look beyond his economics. In particular, Orwell was not a utopian socialist, as were so many of his contemporaries. He believed that socialism had strong tendencies toward centralization and that centralization in turn had strong tendencies toward totalitarianism.

This view was based on two things. The first was Orwell's observation that the Soviet Union, on which so many English socialists placed their hopes, was unmistakably totalitarian. Orwell expressed this opinion on many occasions and in many ways.

Since 1930 I had seen little evidence that the USSR was progressing towards anything that one could truly call Socialism. On the contrary, I was struck by clear signs of its transformation into a hierarchial society, in which the rulers have no more reason to given up their power than any other ruling class.

[*Essays*, Vol. III, p. 405]

Orwell was disgusted with English socialists, because they failed to point out the tyranny which existed in the Soviet Union. In fact, they seemed to him to feel obligated to defend every Soviet action. In Orwell's opinion, these Soviet apologetics were destroying the chances of true socialism ever being established in Great Britain.

It was only *after* the Soviet regime became unmistakably totalitarian that English intellectuals, in large numbers, began to show an interest in it.

[*Essays*, Vol. IV, p. 179]

[N]othing has contributed so much to the corruption of the original idea of Socialism as the belief that Russian is a Socialist country and that every act of its rulers must be excused, if not imitated.

And so for the past ten years I have been convinced that the destruction of the Soviet myth was essential if we wanted a revival of the Socialist movement.

[*Essays*, Vol. III, p. 405]

The other factor which led Orwell to worry about socialist totalitarianism is the obvious fact that central economic planning requires someone to have the power to make and enforce the central economic plan. That person or group of people will have enormous power over their fellow citizens. Such power can be used for vicious political ends, as in Stalin's Russian or in Orwell's Oceania, as well as for the benevolent purposes of economic planning: "[I]t has always been obvious that a planned and centralised society is liable to develop into an oligarchy or a dictatorship" (*Essays*, Vol. IV, p. 163). And Orwell did believe that economic planning was a central feature of the more efficient and just economic system that he envisioned and called socialism.

So Orwell thought that capitalism was unjust and inefficient, but that socialism, the best alternative he could come up with, would lead toward totalitarianism. He expresses this view most starkly in his book review of Frederick von Hayek's *The Road to Serfdom*: "Capitalism leads to dole queues, the scramble for markets, and war. Collectivism leads to concentration camps, leader worship, and war" (*Essays*, Vol. III, p. 119).

Thus Orwell's view of society's future was far more pessimistic than those of his contemporaries, and at a much deeper level. He sees no solution to the dilemma. Although this troubled him greatly, it did not seem to lead him to reexamine either his view of capitalism or of socialism. In economists'

jargon, no equilibrium existed in Orwell's model of the world. Capitalism was unstable in one direction while socialism was unstable in another. Orwell never saw a way out and spent a lifetime living with the belief that civilization was on its last legs. "When one considers how things have gone since 1930 or thereabouts, it is not easy to believe in the survival of civilisation" (*Essays*, Vol. IV, p. 248).

## II. Orwell's Views on Technology

Orwell's views on socialism and totalitarianism didn't win him many friends among intellectuals of his era. His views on technology probably wouldn't win him many friends today, and would probably surprise the many commentators who have claimed that *1984* is an attack on technology.<sup>3</sup> One of Orwell's prime beliefs was that machines have both increased the standard of living of the average man and have relieved him from all sorts of menial drudgery:

From the moment when the machine first made its appearance it was clear to all thinking people that the need for human drudgery, and therefore to a great extent for human inequality, had disappeared. ... And in fact, ... the machine did raise the living standards of the average human being very greatly over a period of about fifty years at the end of the nineteenth and the beginning of the twentieth centuries.

[1984, p. 156]

Thus, Orwell, like many other socialists of his day,<sup>4</sup> thought that improved technology provided the means for ever increasing standards of living.

Surely his most startling view on technology is that all technology should be widely available to the ordinary citizen and not controlled by a central government authority. In a remarkable article called "You and

the Atom Bomb," written in 1945, he applies this line of reasoning even to atomic weapons.

Some months ago, when the bomb was still only a rumour, there was a widespread belief that splitting the atom was merely a problem for the physicists, and that when they had solved it a new and devastating weapon would be within reach of almost everybody. (At any moment, so the rumour went, some lonely lunatic in a laboratory might blow civilisation to smithereens, as easily as touching off a firework.)

Had that been true, the whole trend of history would have been abruptly altered. The distinction between great states and small states would have been wiped out, and the power of the State over the individual would have been greatly weakened....

...[T]hough I have no doubt exceptions can be brought forward, I think the following rule would be found generally true: that in ages in which the dominant weapon is cheap and simple, the common people will have a chance....

Had the atomic bomb turned out to be something as cheap and easily manufactured as a bicycle or an alarm clock, it might well have plunged us back into barbarism, but it might, on the other hand, have meant the end of national sovereignty and of the highly-centralised police state.

[*Essays*, Vol. IV, pp. 7-9]

One can't help wondering what Orwell would have thought about gun control ordinances.

Thus, much of the commentary about Orwell this past year has simply been wrong. The book *1984* is not about television screens and computers, as we have so often been told; it is a book about totalitarianism and socialism. It is a book, not so much written as a prophecy, but as an alarm bell, rung by a man who hardly dared to hope that anyone would answer it. *1984* is a book born of a profoundly pessimistic worldview. I now turn to the world in which Orwell lived and which created this worldview.

<sup>3</sup>See, for example, the essays in Irving Howe (1983).

<sup>4</sup>Oscar Wilde, for example, held this view even more strongly than Orwell. See Orwell's review of *The Soul of Man under Socialism* in *Essays*, Vol. IV., pp. 426-28.

### III. Orwell's Era and Its Mood

Orwell was born in England in 1903 and died in 1950. He lived through one of the most calamitous periods of modern history. As a teenager, he saw World War I; as a young man, he lived through the depression; and in his prime, he experienced World War II. He saw the rise of Bolshevism in Russia; he watched as the promise of communism was betrayed by Stalin. He observed Hitler and the rise of fascism in Europe. He personally fought in the Spanish Civil War against the fascism of Franco. While he was there, he saw how the Soviet-backed Communists fought their anarchist and socialist allies as much as they did their fascist opponents. He foresaw the fall of Spain to fascism (see Orwell, 1980).

The early twentieth century, then, was an era of despair. The generation that fought World War I was often described as the "Lost Generation," partly because so few of them returned, and partly because those who did return were so broken in spirit. Indeed, World War I was widely regarded as the end of civilization. Orwell some years later captured this mood when he wrote:

Many people have remarked nostalgically on the fact that before 1914 you could travel to any country in the world, except perhaps Russia, without a passport. ... Clearly, that is not the kind of social atmosphere that we shall ever see again, and when Sir Osbert Sitwell writes of "before 1914" with open regret, his emotion can hardly be called reactionary. [*Essays*, Vol. IV, p. 443]

Even if the war was not the end of civilization, it certainly was the death knell of classical liberalism, and its optimistic notions of economic freedom and limited government. The Great War was widely regarded as a failure of classical liberalism, and only partly because the Liberal Party was in power when England entered the war. The mere fact of the war itself cast serious doubt on the classical liberal view that free trade and free migration were sufficient to preserve world peace. This view had been a cardinal

tenet of British liberalism in the nineteenth century.

If the Great War was a severe blow to classical liberal politics, the Great Depression was an even more severe blow to classical liberal economics. It was widely believed that the depression was the result of massive market failure, and that the *laissez-faire* policies of the liberals were obsolete. But abandoning liberal economic policies also meant abandoning much of the optimism of liberalism. The idea of an Invisible Hand is itself profoundly optimistic. It says that the common good is being served, even when it is not apparent how. Orwell himself recognized the optimism of the liberals in the following passage: "[T]he Left has inherited from Liberalism certain distinctly questionable beliefs, such as the belief that the truth will prevail and persecution defeats itself, or that man is naturally good and is only corrupted by his environment" (*Essays*, Vol. IV, p. 410).

But all of these notions of economic liberalism went decidedly out of fashion during the early years of Orwell's life. To be sure, socialist arguments against liberalism had been around for some time. But by the middle of Orwell's life, the socialists dominated both the intellectual life and the policymaking process in Great Britain.

This is the sense in which Orwell was truly a child of his era. He shared the pessimism of the Lost Generation. Like his contemporaries, he took for granted that capitalism was dead. This explains the almost offhand nature of many of his comments about economics. "Everyone knew" that capitalism was dead. A detailed defense of that position was not necessary. What was necessary was to find alternatives, no matter how radical, to capitalism. Many people believed that "the Soviet Experiment" was that alternative. But, as we have seen, Orwell had serious reservations about that road to socialism.

### IV. Flaws in Orwell's Economic Vision

It is easy to see how Orwell came to believe that capitalism leads to breadlines and socialism leads to totalitarianism. Both of these ideas seemed to be borne out by ob-

servation of his immediate world. The problem with Orwell's economic analysis is that he accepted the interpretation of the Great Depression that was standard at his time, but which has since undergone serious revision. He also had no understanding of the difficulties of scientific socialism, again a problem which has been widely discussed since his time.

In Orwell's time, the Great Depression was widely regarded as proof of the instability of *laissez-faire* policies. Indeed, this interpretation gave rise to the Keynesian Revolution which prescribed more active intervention by the central government into the management of the economy. But this view has been widely challenged in the intervening years. Milton Friedman and Anna Schwartz's critique (1963) of activist policy included a reinterpretation of the causes of the Great Depression. His view is that, far from being a failure of *laissez-faire*, the depression was a failure of intervention by the Federal Reserve. Now this view is not universally accepted, but then again, neither is the old view universally accepted. Today at least there is some challenge or alternative to the view which Orwell and his contemporaries took for granted. One must at least consider the possibility that ill-conceived intervention contributed to the depression.

But it would be asking a great deal of Orwell to expect him to anticipate this neo-classical critique. A greater failure is that he had no understanding of the difficulties of "scientific socialism," as it was called. A great debate raged during the 1920's and 1930's on the possibility of industrial socialism.<sup>5</sup> One school of thought, led by Oskar Lange and Abba Lerner, argued that central planning could be done efficiently, provided that the planners used marginal cost shadow pricing to simulate market pricing. The op-

posing school of thought, led by Ludwig von Mises and the Austrians, argued that the advocates of marginal cost pricing were overlooking the essential difficulty of central planning. The Austrians argued that in the absence of a market process, there is no reliable way to discover marginal costs and efficient levels of output and efficient methods of production. For a review of this debate, see Trygve Hoff (1981).

This criticism has been repeated in many forms for many problems. Economists often criticize both "industrial policy" and "comparable worth" on these grounds. How can government industrial policy choose the potential "winners and losers" in the world marketplace in the absence of the actual competitive process? How can the judiciary determine jobs of comparable value without reference to market wages? The fact that these proposals continue to emerge and that some economists support them demonstrates that opinion on these matters is by no means uniform, even among economists. However, it is naive to assume, as Orwell seems to have, that planning an economy is a straightforward extension of the exercise of planning a family shopping list. He seems to have thought that if one activity requires planning in order to be successful, then surely the other does as well.

The essential missing link in Orwell's understanding of economics is that he had no notion of the price system as a coordination mechanism. Price changes convey information to all of the diverse economic agents, and this information allows them to coordinate their plans. And based on this minimal information, the agents make their own plans, and take their own actions. The adjustments of the price system make it possible for all of those plans to be coordinated. This is the basic meaning of the concept of equilibrium. That is why central planning is not simply individual planning on a larger scale. The central planner must coordinate the plans of the many individuals, a job which the market economy leaves to the Invisible Hand.

Orwell seemed to have no appreciation of the magnitude of the coordination problem that the price system attempts to solve. In short, Orwell had no notion of what is

<sup>5</sup>The following passage suggests that Orwell had some knowledge of this debate: "I am well aware that it is now the fashion to deny that Socialism has anything to do with equality. In every country in the world a huge tribe of party-hacks and sleek little professors are busy 'proving' that Socialism means no more than a planned state-capitalism with the grabmotive left intact" (1980, p. 104).

sometimes called spontaneous order in economics. If there is no explicit planner, there must be no plan. And without a plan, there must be no order.

#### V. Conclusion

These then are the economic ideas which led Orwell to believe that capitalism was doomed: a particular interpretation of the Great Depression, and no concept of spontaneous order or market process. It is interesting to speculate about how Orwell's thought might have developed had he not died prematurely in 1950 at the age of 47. Had he observed the problems of British socialism and of the major centrally planned economies, he may have revised his estimate of the ease and desirability of central economic planning. He had always had a distinct mistrust of the central political power inherent in socialism. Perhaps if he had observed some of the economic difficulties of socialism, he may have rejected it outright, rather than trying to reform it. Of course, he may have just hated socialism along with capitalism. If so, this would have added to the profound

pessimism which he felt during all of his short life. He expressed it rather poetically in a letter to his friend Arthur Koestler in the spring of 1946, "Each winter I find it harder and harder to believe that spring will actually come" (*Essays*, Vol. IV, p. 127).

#### REFERENCES

- Friedman, Milton and Schwartz, Anna J., *A Monetary History of the United States, 1867-1960*, NBER Studies in Business Cycles, Vol. 12, Princeton: Princeton University Press, 1963.
- Hoff, Trygve J. B., *Economic Calculation in the Socialist Society*, Indianapolis: Liberty Press, 1981.
- Howe, Irving, *1984 Revisited*, New York: Harper & Row, 1983.
- Orwell, George, *Homage to Catalonia*, New York: Harcourt, Brace, Jovanovich, 1980.
- \_\_\_\_\_, *1984*, New York: New America Library, 1983.
- Orwell, Sonia and Argus, Ian, *The Collected Essays, Journalism and Letters of George Orwell*, Vols. I-IV, New York: Harcourt, Brace, Jovanovich, 1968.

## Hamlet without the Prince: Cambridge Macroeconomics without Money

By J. A. KREGEL\*

Keynes' *General Theory* was exclusively concerned with a monetary economy in which changing beliefs about the future influence the quantity of employment. Yet money plays no more than a perfunctory role in the Cambridge theories of growth, capital, and distribution developed after Keynes. This essay attempts to explain this paradox with reference to the relation between Keynes' monetary revolution and the value theory revolution which simultaneously occurred in Cambridge in the 1930's.

### I. The Instability of Credit

The *General Theory* has often been described as provoked by the Slump, yet F. Vicarelli (1984) argues that it also represents the theoretical formulation of reflection on the nature of capitalism begun in the 1920's. Hayek describes this period: "We all held similar views...more fully elaborated by R. G. Hawtrey who was all the time talking about the inherent 'instability of credit'" (in Milton Friedman, 1969, p. 88n). Indeed R. G. Hawtrey's 1913 *Good and Bad Trade* foreshadowed a monetary theory of effective demand:

[T]he manufacturer's efforts in producing...goods depends upon there being an effective demand for them.... It is only because the dealer anticipates... this effective demand...that he gives the manufacturer the order.... The manufacturer...accepting the order, and the banker...discounting the bill,

are both endorsing the opinion of the dealer. The whole transaction is based ultimately on an expectation of a future demand, which must be more or less speculative. [p. 78]

The impact of money on the level of activity inspired not only Hayek and Keynes, but Robertson, Schumpeter, Pigou, Cassel, and others. These economists, to whom Keynes addressed his 1923 *Tract on Monetary Reform*, all implicitly accepted the quantity theory of money; open rejection of this common analytical framework in the 1930 *Treatise on Money* made communication with his contemporaries difficult, and Keynes had to look elsewhere for sympathetic criticism.

### II. Cambridge Value Theory in the 1930's

At the same time, a group of young Cambridge economists, Kahn and Joan and Austin Robinson, were extending Sraffa's 1926 criticisms of Marshall's theory of value to produce the "Imperfect Competition Revolution." It was among these economists, and others involved in developing imperfect competition such as Harrod, Kaldor, and Kalecki, that Keynes' ideas created interest. They did not, however, have first-hand experience of the earlier "monetary" debates, indeed, for many skepticism of the quantity theory (see R. F. Kahn, 1984, p. 52) deterred them from the study of monetary factors which had inspired Keynes' generation. Thus the Cambridge economists who formed the "Circus," and others such as Harrod, who played a central role in discussing and elaborating the *General Theory*, were all involved in the value theory revolution before they joined Keynes' monetary revolution. It is not surprising that they should have perceived a relationship, although Keynes believed that developments

<sup>†</sup>Discussants: Donald Harris, Stanford University; Mark Kuperberg, Swarthmore College.

\*Professor of Economics, University of Groningen, The Netherlands.

in value theory were independent of his own pursuits.

The *Treatise* replaced the equation of exchange as determinant of the price level with the "fundamental equations" which combined Marshall's short-period supply and demand factors in a sort of reduced-form equation representing the composition of expenditure relative to the composition of output. Money retained only indirect influence on prices via the influence of the rate of interest on the divergence of saving and investment producing "windfall profits." Here Keynes followed Wicksell and called "natural" the interest rate that produced price equilibrium. Keynes' new interlocutors were to draw on their value theory expertise to provide two crucial elements for the transformation of the "fundamental equations" into the analytical framework of the *General Theory*: Kahn's analysis of short-period supply and Sraffa's conception of commodity rates of interest.

#### A. Kahn's Short-Period Analysis

It was Mr Kahn [in his multiplier article] who first attacked the relation of the general level of prices to wages in the same way as that in which that of particular prices had always been handled, namely as a problem of demand and supply in the short period rather than as a result to be derived from monetary factors.

[Keynes, 1939, in 1973, p. 400, fn. 1]

Kahn also convinced Keynes that the price of investment goods should be handled in an analogous fashion, thus separating the combined supply and demand aspects of the fundamental equations which opened the way to an aggregate demand function representing consumption expenditures (from wages) and investment expenditures (from profits) and an aggregate supply function to determine the "aggregate" price level as an application of his value theory investigations into "short-period supply" to Keynes' monetary studies. While others, such as Hicks, were applying value theory to the demand for money, in Cambridge it was being applied to the "fundamental equations."

Indeed, Kahn (p. 99) considers his greatest contribution to Keynes' theory as the demonstration that investment generates the savings required to finance it, which convinced Keynes to adopt savings-investment equality. This eliminated the second term of the fundamental equations and facilitated application of aggregate supply and demand analysis, but it also eliminated money from even an indirect role via interest rates on prices. Sraffa was to provide the new role for money and the rate of interest.

#### B. Sraffa's Commodity Rates of Interest

In his 1932 review of Hayek's *Prices and Production*, Sraffa formulated commodity rates of interest to criticize Hayek's use of Wicksellian "natural" rates of interest (which had equated saving and investment in the *Treatise*). Hayek's theory suggested that if the presence of money was "neutral," the natural rate of interest would equate investment to full employment saving and Say's Law would necessarily prevail.

Hayek's "neutrality" required the money rate of interest equal the "equilibrium" (natural) interest rate so that money savings should buy the same amount of producers' goods as if the supply and demand for capital met in "their natural form." Sraffa pointed out that natural rates exist implicitly for every commodity, and explicitly whenever there is a forward market. These rates will be uniform in an equilibrium in which "the spot and forward price coincide" for each commodity. But when savings, whether in money or natural form, are positive there may be as many natural rates as there are commodities for:

[U]nder free competition this divergence of rates is as essential to... transition as is the divergence of prices from the costs of production; it is, in fact, another aspect of the same thing.... [This] applies as much to an increase in saving, which Dr. Hayek regards as equivalent to a shift in demand from consumers' to producers' goods, as to changes in the demand for or the supply of any other commodity.

[1932, pp. 50-51]



Since it is the competitive price adjustment process, not the nonneutrality of money which causes divergence of individual commodity rates and these from the money rate, neutrality cannot be defined in the conditions Hayek proposed.

Sraffa's criticism suggested that the influence of money was not via interest rate effects on saving and investment, but via the divergence of commodity rates which was just "another aspect" of the divergence of demand prices from supply prices leading to production decisions for consumption or investment goods. The rate of interest provided a parallel representation at the level of individual production decisions for the divergence of aggregate demand and supply prices which was emerging from the elaboration of the fundamental equations, but one in which money was clearly central to the determination of the expenditure decisions which brought changes in production and employment.

### III. Interest Rate Parity and Liquidity Preference

In Keynes' monetary economy, money offered an alternative to investment in other durables; it thus required a comparable definition of its rate of interest in terms of spot and forward prices. Keynes thus defined the money rate of interest as "nothing more than the percentage excess of a sum of money contracted for forward delivery, e.g. a year hence, over what we may call the 'spot' or cash price of the sum thus contracted for forward delivery" (1973, p. 222). This change in the way money entered Keynes' analysis allowed an analogy with the *Tract's* "interest rate theorem" (see my 1982 article) to explain decisions to take positions in durables (including money) since every durable has spot and forward prices in terms of itself as numeraire giving its "commodity" or "own-rate of own interest": "just as there are differing commodity-rates of interest at any time, so also exchange dealers are familiar with the fact that the rate of interest is not even the same in terms of two different moneys..." (Keynes, 1973, p. 224).

In the exchange market, equilibrium is achieved when the forward discount or pre-

mium reaches equality with the interest differential. Just as the forward discount in the *Tract* measured the market's "preference" for holding that currency, Keynes now argued that the market's preference for holding money was also measured by the discount of future over spot money in terms of money: the rate of interest. Equilibrium was thus defined by the equalization of the relative advantages of taking positions in durable assets, that is, when all the own-rates evaluated in money were equal to the own-rate on money.

In this way Keynes gave expression to the way "changing views about the future are capable of influencing the quantity of employment and not merely its direction" (1973, p. vii), for every decision to purchase (invest in) a durable at prevailing spot prices depends on expectations of future conditions as expressed in future prices: the relation of spot to forward prices or supply and demand prices determines rates of return on investment in durable goods or their marginal efficiencies, while liquidity preference sets the spot and forward price of money, the rate of interest. Marginal efficiencies and liquidity preference thus reflect views of the future or the state of general expectation. The composition of asset holdings and overall expenditure decisions were thus directly influenced by changing views of the future. Since each individual had to reach his or her own view on the relation between present and future prices to determine the investment and expenditure strategy which maximized expected rates of return, decisions to buy investment goods or to hold money would be characterized by diversity of view as to expected rates. Equilibrium would be established when market prices for any activity produced a balance of divergent expectations. In such conditions, as Shackle has stressed, actual events may disappoint every agent's expectations, forcing frequent reconsideration of position and making decisions to invest liable to constant fluctuation. It was only possible to discuss equilibrium on the assumption of a given state of general expectation in which an increase in investment, or shifts between activities, eliminates discrepancies in own-rates evaluated in money

(marginal efficiencies) by means of adjustment in spot prices, current production (i.e., net investment), or the degree of liquidity, depending on the type of market and type of activity (see M. Tonveronachi, 1983, pp. 167-68).

Changes in expected future conditions thus influence divergences in marginal efficiencies which reflect differences between costs of production and prices which initiate expenditure decisions and affect both future supply and prices. Adjustment continues until marginal efficiencies and the money rate of interest are brought into equilibrium. Equilibrium could thus be represented in the aggregate in terms of effective demand equating aggregate supply price or on the individual level as uniformity of own-rates given by spot and forward prices of durables.

Further analysis of the nature of money and the behavior of liquidity preference was necessary to determine whether the equilibrium thus achieved was durable at less than full employment, for if money rates could be brought to a sufficiently low level to lead individuals to spend all of their income on current production, then full employment was the only stable position. Keynes suggested that since interest rates represented expectations of future rates, such a policy would only succeed if individuals could be convinced that reductions in money rates would not be reversed. His simple formula showing how capital loss offsets yield for smaller rises in interest the lower the prevailing rate (1973, p. 202) suggested that healthy skepticism (reinforced by the appearance of "semi-inflation" if money wages progressed more rapidly than productivity as expenditure increased) would lead rational investors to increase liquidity preference, causing the money rate to be the one to "decline most slowly as the stock of assets in general increases" (1973, p. 229), producing equilibrium (and confirming the skeptics' opinions) before full employment is reached.

But orthodox theory had argued that even if increased hoarding reduced the demand for goods, this would only change the direction, not the amount, of employment if money were produced by labor. Keynes' "essential properties of money" rather than the

liquidity trap are meant to meet this point; if the elasticities of production and substitution of money are negligible then the demand for money is not a demand for goods and money provides a "sink" for purchasing power.

#### IV. Own-Rates and Aggregate Supply and Demand

Discussion of a monetary economy characterized by money bearing these essential properties formed the introduction to early drafts of the *General Theory*, reflecting Keynes' announcement of a "Monetary Theory of Production" in the Spiethoff *Festschrift*, but was to be replaced by the short-period equality of aggregate supply and demand price representing the principle of effective demand. Money appears as an integral part of the discussion of the determinants of investment, reflecting Keynes' move towards the identification of investment as the *causa causans* in determining output and employment; while money loses pride of place, it gained in importance, providing the very basis for the explanation of the inherent instability of investment in a capitalist economy, and the explanation for persistent unemployment equilibrium by means of the "essential properties." The final title announces a general theory in which "Employment" is determined by "Interest and Money."

Thus the two influences from the value theory revolution provided parallel frameworks for presentation of Keynes' revolution in monetary theory; the "Marshallian" analysis in terms of aggregate short-period supply and demand prices took on the more visible role in chapter 3, representing the changed role of money as the basic determinant of individual investment decisions determining asset prices reflecting liquidity preference, and explaining unemployment equilibrium.

Hicks was clearly more impressed by the aggregate supply and demand framework, but proposed his own Walrasian basis by aggregating a general equilibrium system into "bundles" representing three "market" equilibria in which equality of the natural and market rate of interest replaces that of "aggregate" supply and demand prices of the

principle of effective demand. This formulation placed money within the Walrasian framework and replaced Keynes' discussions of money and investment decisions with a constraint on the demand for money so as to produce a horizontal "Keynesian" range to the *LM* curve where a stable interest rate (the only price) also implied fixed wages and prices. Keynes vigorously denied such a relation and instead suggested that: "the difference between myself and the classicals lies in the fact that they regard the rate of interest as a non-monetary phenomenon" (1973b, p. 80). Keynes felt that his analysis of the effect of money and interest on investment had been overlooked because readers had placed undue emphasis on the difference between expected and actual demand. He repeats the point made to Hicks in a series of articles published in 1937, arguing that in his theory money was a "real" factor which could affect relative money prices and outputs in the long period as well as the short, while in difference to the theories of Wicksell and Hayek, the rate of interest was a purely monetary factor, independent of any real or natural forces:

Put shortly, the orthodox theory maintains that the forces which determine the common value of the marginal efficiency of various assets are independent of money, which has...no autonomous influence, and that prices move until the marginal efficiency of money, i.e., the rate of interest, falls into line with the common value of the marginal efficiency of other assets as determined by other forces. My theory, on the other hand maintains that this is a special case and that...the opposite is true, namely that the marginal efficiency of money is determined by forces partly appropriate to itself, and that prices move until the marginal efficiency of other assets fall into line with the rate of interest.

[1973b, p. 103]

## V. Extensions of the General Theory

Thus there were two possible lines of development: extension of the short-period

Marshallian (or Hicks' Walrasian) aggregate supply and demand framework to analyze factors such as capital accumulation, which were traditionally treated as "long period" problems, or analysis of a monetary economy where money is a determinant of the investment decision within the own-rate framework. Most economists, including Keynes' younger colleagues, pursued elaboration of long-period analysis and Kahn, Robinson, Kaldor, Harrod, among others, built on the aggregate version of Marshall's short-period to tackle the long-period problems of capital accumulation and distribution. Thus the Harrod-Domar growth models spawned multiplier-accelerator models, and the analysis of capital accumulation produced the Cambridge theories of aggregate income distribution. It is characteristic of these latter theories that investment is exogenous, eliminating the need to discuss the monetary elements which Keynes had used in the *General Theory* to explain investment decisions. While investment and expectations play a crucial role in the post-Keynes Cambridge theories, the fact that they were considered exogenous made analysis of the monetary factors Keynes considered crucial to their determination unnecessary. Of little importance to the formulation of short-period aggregate supply and demand, monetary factors and Keynes' concerns for cyclical instability had even less importance in the extension of these constructions to stable long-period equilibria.

For economists who preferred Hicks' Walrasian version, the analysis of long-period problems such as capital accumulation suggested flexibility of wages and prices, and variation in the rate of interest adjusting capital intensity to produce full employment. It was thus the aggregate supply and demand version of Keynes' theory, incorporated into traditional theory as a temporary general equilibrium with fixed prices, that failed transition to long-period conditions of flexible prices and production coefficients; the Cambridge objections to these formulations, made in Marshallian terms, had little impact. Indeed, in the Cambridge capital theory debates it was necessary to abandon Marshall and resort to Sraffa's theory of prices to

demonstrate the unlikely conditions required for the traditional results, but this left Keynes' theory still characterized by *ad hoc* short-period rigidity.

Joan Robinson, after a long career forging the long-period extension of the Marshallian short-period version of Keynes, finally rejected this approach in favor of the study of "historical" time, but has recently suggested (1978) that post-Keynesian analysis should be extended to reconcile Keynes and Sraffa. This would imply integrating Keynes' essential monetary relationships with Sraffa's. Keynes' extension of Sraffa's commodity rate analysis to a monetary economy was clearly an attempt to respond to Hayek by showing just exactly why money was not neutral. Just as Sraffa went on to develop a similar framework to specify the essential logical relations between distributive variables and relative prices, Keynes' extension of the interest rate parity theorem can be considered as the identification of the logical relations between the money rate of interest and the level of output and employment in a monetary economy. While Sraffa's relations imply no particular causal relationships, Keynes' system explicitly proposes money as the real or causal variable constraining a capitalist or monetary production system. It is the recognition of this role of money that has been absent from Cambridge, and all other, macroeconomics after Keynes.

Recently, there has been a renewal of interest in Keynes' monetary analysis. For example, Paul Davidson (1972) analyzes investment in terms of spot and future prices, drawing on the monetary detail of the *Treatise* to support liquidity preference, while Hyman Minsky (1975) has linked the role of aggregate expenditure to the determination of the relative prices of consumption goods and capital assets to identify the influence of the financial system on investment cycles. Luigi Pasinetti's (1981) discussion of the relation between natural rates and the social relations producing money rates of interest builds on similar factors. I (1982, 1983) and E. J. Nell (1983) attempt to draw direct links between the concepts of own-rates in Keynes and Sraffa.

From Hicks's reformulation of Keynes' supply and demand analysis in Walrasian terms to Don Patinkin's recent emphasis on its Marshallian origin, the profession (with the exception of D. Dillard, 1954, and more recently A. Barrère, 1985) has considered the principle of effective demand independently of Keynes' explicitly announced intention to analyze a monetary production economy and his claim that the distinctive aspect of his theory was to be found in the essential properties of interest and money which made the latter a real factor and the former a monetary factor. Recognition of this distinction as expressed in Keynes' own-rate framework may yet produce the monetary revolution in economic theory promised by Cambridge economics after Keynes.

## REFERENCES

- Barrère, A., *Keynes Aujourd'hui*, Paris: Editions Economica, 1985.
- Davidson, P., *Money and the Real World*, London: Macmillan, 1972.
- Dillard, D., "The Theory of a Monetary Economy," in K. Kurihara, ed., *Post-Keynesian Economics*, London: Allen and Unwin, 1955, ch. 1.
- Friedman, M., *The Optimum Quantity of Money*, London: Macmillan, 1969.
- Hawtrey, R. G., *Good and Bad Trade*, London: Constable, 1913.
- Kahn, R. F., *The Making of Keynes' General Theory*, Cambridge: Cambridge University Press, 1984.
- Keynes, J. M., (1973a) *Collected Writings*, Vol. VII: *The General Theory of Employment, Interest and Money*, London: Macmillan, 1973.
- \_\_\_\_\_, (1973b) Vol. XIV: *The General Theory and After, Part II*, London: Macmillan, 1973.
- Kregel, J. A., "Money, Expectations and Relative prices in Keynes' Monetary Equilibrium," *Economie Appliquée*, No. 3, 1982, 35, 449-65.
- Minsky, H. P., *John Maynard Keynes*, London: Macmillan, 1975.
- Nell, E. J., "Keynes after Sraffa: The Essen-

- tial Properties of Keynes's Theory of Interest and Money," in J. A. Kregel, ed., *Distribution, Effective Demand and International Economic Relations*, London: Macmillan, 1983, ch. 2c.
- Pasinetti, L. L., *Structural Change and Economic Growth*, Cambridge: Cambridge University Press, 1981.
- Robinson, Joan, "Keynes and Ricardo," *Journal of Post Keynesian Economics*, Fall 1978, 1, 12-18.
- Sraffa, P., "Dr Hayek on Money and Capital," *Economic Journal*, June 1932, 42, 42-53.
- Tonveronachi, M., J. M. Keynes. *Dall'Instabilità Ciclica all'Equilibrio di Sottoccupazione*, Rome: Nuova Italia, 1983.
- Vicarelli, F., *Keynes: The Instability of Capitalism*, London: Macmillan, 1984.

# Cambridge Price Theory: Special Model or General Theory of Value?

By BERTRAM SCHEFOLD\*

Keynes and Marshall liked to emphasize the continuity of their thought with the classical Ricardian tradition. Although this affinity concerned aspects of their theories which were, as theories of value, essentially neoclassical, the followers of Keynes—Joan Robinson in particular—began to return to a classical analysis of prices after the foundations of the neoclassical theory of output had been shaken by Keynes. The ground for this had been prepared by the debate about increasing returns (in which Sraffa's contribution was prominent), the recurrent discussions about capital theory and the study of classical authors. The appearance of Sraffa's *Production of Commodities by Means of Commodities* marked a second revolution of Cambridge economics in this century. Neoclassical economists have reacted to Sraffa's departure much in the same way as they had reacted to Keynes': by trying to integrate the new ideas as special cases of the old doctrine. However, the Cambridge price theory is, contrary to Frank Hahn (1982), not a model within the neoclassical universe, but a reformulation of the classical theory.

The general equilibrium of the long period is by definition an equilibrium in all markets, including the capital market, so that it must be characterized by a uniform rate of profit. If homogeneous commodities are produced by single-product industries, Sraffa's price equations are

$$(1) \quad (1 + r)Ap + wl = p,$$

where the coefficient  $a_{ij}$  of the input-output matrix  $A$  denotes the amount of commodity

$j$  and  $l_i$  the amount of labor required to produce one unit of output in industry  $i$ , where  $p$  is the price vector,  $w$  is the wage rate, and  $r$  the rate of profit.

The socially necessary technique and the level and composition of output are considered as given (constant returns are not assumed). The determination of the composition of output through individual preferences is formally not ruled out, but other explanations are available. The theory of distribution is treated as a separate issue. The function of prices is to provide a measure of normal costs against which the profitability of alternative technologies and future investments can be estimated.

Prices of production are such that the economic system can reproduce itself in its actual state of growth or decline if the underlying conditions remain unchanged. If they do change, the system will enter a new long-period position with new prices of production. The prices of production act as centers of gravitation for market prices, but this does not mean that market prices will necessarily ever get very close to prices of production; the theory only denies that the discrepancy between market prices and prices of production can become permanently large, involving considerable losses or surplus profits, while the fundamental conditions remain the same. The cause for the gravitation of market prices is the tendency for the equalization of rates of profit in different industries. But the process may take diverse forms. Apart from changes in prices themselves, transfers of funds, investment, capital losses and the fixation of surplus profits as rents through the establishment of new property rights are all important. New institutions which confirm the tendency are invented as often as institutions which upset it. The classics were therefore not very interested in the many conceivable models that might describe processes of gravitation.

\*FB 2, J. W. Goethe-Universität, Frankfurt/M., Germany. I thank V. Caspari, G. Duménil, G. Englmann, J. Eatwell, P. Garegnani, D. Lévy, and others for helpful discussions.

In modern times, adjustment of capacity utilization at unchanged prices seem often to be more important than changes of market prices for the adaptation of supply and demand to temporary disturbances. Prices and markups reflect the conditions that would prevail if production were steady at normal capacity utilization. Up to a point, changes in demand can then be met through adjustments in stocks, orders, and levels of operation without changes of prices so that market prices can actually be equal to prices of production although profitability fluctuates as with changing market prices.

At any rate, prices of production are conceptual magnitudes defined for the analysis of an economic system in a given state of accumulation. They are necessary as measures for the distribution of the surplus, for the quantity of capital, and for the analysis of different forms of technical progress.

The concept of equilibrium and of stability is different in *modern* neoclassical theory. If the economy is not in a stationary or atemporal equilibrium (with which we are not concerned here), one modern neoclassical method is that of *temporary* equilibrium. In such an economy, there is no trading in forward markets. Prices will be such that, in a given state of expectations, markets clear and existing stocks can be used to produce goods according to the anticipated state of future demand. Such a model deals only with market prices in the short run. Moreover, the stability of this kind of equilibrium is problematical, because small accidental deviations may alter the expectations which define it.

Alternatively, the atemporal equilibrium model is extended to an *intertemporal* one with a finite or infinite horizon; there are then present markets for future deliveries, in terms of discounted prices. Contracts can be contingent on changing future states of nature. It is therefore possible to represent the classical problem of growth with technical change, but not as a process going on through time. Instead, that set of prices is chosen which will coordinate demand and supply in present and future markets simultaneously, given the expected changes in the structure of production and consumption. Positive prices are exchange ratios which clear pre-

sent and future markets. At the same time, they define the factor income by households and differential profits (obtained by firms under advantageous conditions and distributed to households). But, to the extent that production takes time, inputs to a given process are not valued at the same prices as outputs. Each good exists at a particular point in time and is to be valued at the prices which bring contracts for deliveries for that moment of time into equilibrium.

The own-rate of interest of a commodity can be defined as the maximum excess (in percent) that can be obtained of a commodity at the end of a period over one unit of the commodity invested in the beginning. In an intertemporal equilibrium with discounted prices, this must be equal to the ratio of the price at the beginning divided by the price at the end.

It is well known that there may be as many own-rates of interest as there are commodities in a neoclassical intertemporal equilibrium. The own-rate of interest is equal to something that looks like the unique rate of profit in the classical system if a commodity or commodity basket is taken as the numeraire: let  $a_i$  be a vector of inputs (including factors) in period  $t$  to a process  $i$  which yields outputs  $b_i$  at the beginning of period  $t+1$  in an intertemporal equilibrium with price vectors  $p_t$  and  $p_{t+1}$  for the beginning and the end of period  $t$ . No net profits will be made in a neoclassical intertemporal equilibrium at *discounted* prices so that  $a_i p_t = b_i p_{t+1}$  for all processes  $i$ . However, we can express prices in each period in baskets of goods given by some vector  $d$ —this includes the case where  $d$  represents a producible monetary commodity. These *undiscounted* prices in terms of numeraire  $d$  may be denoted by  $p_d$ ; we have  $p_{d,t} = p_t / dp_t$  so that we may calculate a rate of return  $r_d$  for process  $i$  in terms of commodity basket  $d$  and obtain

$$(2) \quad 1 + r_d = b_i p_{d,t+1} / a_i p_{d,t} \\ = (b_i p_{t+1} / dp_{t+1}) / (a_i p_t / dp_t) = dp_t / dp_{t+1}.$$

If we reckon in terms of undiscounted prices, all processes are therefore equally profitable but the rate of return depends on the commodity chosen as the numeraire; it is equal

to the own-rate of interest of the standard of prices in each period.

All own-rates of interest are necessarily equal only if the vectors of relative prices are proportional in different periods as in the classical model. As a matter of fact, Sraffa was the first to provide a skeptical account of the multiplicity of own-rates of interest in intertemporal equilibrium in his review of Hayek (1932). Hayek had been looking for a "natural" rate of interest which might guide monetary policy; it turned out that his intertemporal equilibrium contained many. Ever since, neoclassical theorists have been pursuing the impossible task of finding "the" rate of interest in intertemporal equilibrium as a monetary phenomenon while preserving the neutrality of money. One approach is based on the empirical lack of forward markets and leads back to a temporary equilibrium in which there are forward markets for only one monetary commodity which yields the rate of interest. This again raises the question mentioned above of what stability such an equilibrium might have if expectations are not derived from fundamental data such as innate preferences. Moreover, it is not clear how this would have to be extended to an analysis of accumulation going on through time in a sequential process. In particular, the point of view of intertemporal equilibrium suggests that the rate of interest depends on long-term contracts while the post-Keynesians in Cambridge have emphasized the possibility of a dependence of the rate of interest on short-run phenomena that may be independent of the more permanent influences on the rate of profit in industry.

The strangest aspect of the modern intertemporal equilibrium theory concerns the treatment of initial endowments; they are treated as given so that in general some will be in excess supply and will receive zero prices if there are limited possibilities of substitution. The rates of return are therefore not uniform in that the stocks in excess supply yield no return at all. Hence, the fundamental condition of a long-period equilibrium is violated. Moreover, according to a distinct second argument, the stocks must once have been acquired in order to use them

profitably so that, if they now turn out to be superfluous, we are faced with a short-period problem of adaptation and the deceived expectations should—but do not in the model—influence the future course of events. This is not a mathematical but a conceptual inconsistency. The model is often defended on the grounds that the entire future allocation over the given time horizon is settled afresh by means of forward trading. Against this, one may not only doubt the empirical validity of the assumptions but may also ask why this foresight did not exist yesterday to prevent the wrong stocks from accumulating. The intertemporal model like the temporary one, does not therefore, represent a "true" equilibrium for all markets, since it fails to take into account the heritage of stocks from the past with their associated expectations, in an analysis of what should be called a disequilibrium situation. The intertemporal model is not wrong because it fails to formalize expectations explicitly, for what matters is what generates expectations. It is a hybrid mixture of long-period equilibrium with uniform rates of return (but they are in general not uniform for endowments) and short-period adaptations to normal conditions (but then the disturbances should be treated as such even if expectations are only implicit). Finally, no distinction is made between the reaction of investment to temporary surpluses or shortages in particular industries on the one hand, and changes of aggregate effective demand on the other.

Unfortunately, classical and neoclassical models tend increasingly not to be compared as fully developed economic theories of value, money, employment, growth, and development; the tendency is to confront directly only the theories of value. If these are not (as, for example, by Schumpeter) seen in their contexts, the comparison may be reduced to a confrontation of the formal properties of models. It may then appear that the classical theory as represented by Sraffa's equations is only a special case of the intertemporal general equilibrium. The argument is that the classical model is characterized by a uniform rate of profit, with undiscounted prices of outputs and inputs being the same. Price vectors in the neoclassical model are



proportional if there is balanced growth. Hence it is concluded that the classical model can only be concerned with the situation in which endowments and preferences happen to be such that the economy can start and follow a process of balanced growth.

As a matter of fact, there are important contextual differences between classical and neoclassical theories even if balanced growth is assumed or attained. Joan Robinson was concerned with the analysis of "Golden Ages" where she supposed stocks already to have been inherited in a balanced composition because those stocks had been accumulated with a view to sustain future growth. She then examined the conditions, in particular in relation to effective demand, which would sustain expectations consistent with growth at various levels of employment.

Now it should be clear that *balanced* growth plays a special role in *neoclassical* theory because, with constant returns and constant preferences, the economy will be close to a balanced growth path (turnpike theorem) anyway and, more importantly, because different scarcities of different capital goods will not be present and cause unequal rates of return in terms of the same standard in the "steady state." But the problem of *unbalanced* growth or of structural change had been one of the main concerns of *classical* theory from its early beginnings.

Ricardo's vision of the extension of cultivation to inferior lands with the growth of population in the process of accumulation involved such a structural shift, and he analyzed it by comparing what we now call systems of prices of production in subsequent long-period positions, where the rate of profit could be uniform in each, because the growth of receipts over costs on better lands was absorbed by rents. Changes in the composition of output due to the influence of varying income elasticities with the growth of real incomes can be dealt with by means of the same method. Finally, technical progress causes profit differentials which express the fact that two systems, corresponding to an "old" and a "new" technique, coexist for some time. The differential will be the more lasting the greater is the market power of firms, but imperfections of competition have

to be treated as a separate matter, and the comparison of long-period positions remains the essential step of the analysis.

There are special cases of structural change which are analyzed best by admitting a slow, continuous shift of the input-output coefficients. The outstanding examples are the effects of technical progress as they appear at higher levels of aggregation in Leontief systems, or the change in the composition of output that accompanies rises of real incomes and is described by Engel curves. If the coefficients of  $A$  and  $l$  depend on time and follow some foreseeable pattern, prices will do the same. The unique rate of profit can still be defined if the movement is slow relatively to the movement of production; one then has, in obvious notation

$$(3) \quad (1 + r_t)A(t)p_t + w_t l(t) = p_t.$$

The own-rate of interest as defined above is for each commodity  $i$  equal to the rate of profit. But it is in general not equal to  $p_{t,i}/p_{t+1,i}$ , because we are here dealing with undiscounted prices.

Key problems of the theory of unbalanced growth arise if the process of change is accelerated so that, for instance, the consumption of some commodities shrinks faster than capacity for their production does because of wear and tear. Such an increase of excess capacities may have implications for aggregate effective demand and destroy the stability of the long-period position. The discontinuity must then again be analyzed in terms of comparisons of equilibria.

If structural change is fast, one might like to introduce shifts of relative prices by allowing relative prices at the beginning of the period to be different from those at the end. But is easy to see that effects of changes in distribution are then not easily distinguished from effects of changes in the standard of prices. To make a strong case, suppose we had  $A(t), l(t)$  depend on time  $t$ , with  $t = \dots -1, 0, 1, \dots$  stretching indefinitely back into the past and into the future. It is assumed that markets clear at positive prices (no excess stocks) in each period and that relative prices fulfill the infinite series of

equations, which defines a Golden Age with structural change:

$$(4) \dots, (1 + r_t)A(t)p_t + w_t l(t) = p_{t+1}, \dots$$

Under suitable assumptions prices are determined uniquely. If, for instance, productivity grows at rate  $g$  so that  $l(t) = l_0(1 + g)^{-t}$ , if  $A(t) = A$  and  $r_t = r$  are constant, and if the rise of the wage rate compensates productivity gains  $w_t = w_0(1 + g)^t$ , one obtains constant prices and  $r$  is the rate of profit as in a Sraffa system. If  $w_t = w_0$  is constant, prices fall with the rate of productivity  $g$ , but each own-rate of interest exceeds both  $r$  and  $g$  and the deflation implies that the rate of profit exceeds  $r$ . Conversely, wage rates rising faster than productivity lead to an inflation which will let profits appear higher than they really are (if the wage rate rises at rate  $h$ ,  $p_{t+1} = ((1 + h)/(1 + g))p_t$  and the rate of profit is  $(1 + g)(1 + r)/(1 + h)$  at constant prices).

This lack of transparency indicates that the traditional method of analyzing structural change by means of comparisons of equilibria is preferable. The method of production in (3), reached after each gain in productivity, is considered as socially necessary; market prices will adjust to the prices of production so defined with a time lag. But if one insists on using an "intertemporal" system as in (4), with relative prices changing, for example, because of different rates of growth of the productivity of labor in different sectors, the equilibrium is not determined and little can be said about distribution (i.e., the real wage and the rate of profit), before the movement of the wage rate has been specified and the system is reformulated in terms of equations (3).

In spite of the familiar analogy with the compensation of the differences in the rates of interest between two currencies through differences in the spot and forward exchange rates, it is not clear whether an analysis of production in terms of an intertemporal price system could be of much relevance even apart from the lack of transparency. For it involves on the one hand, like an analysis of market prices, quick changes in relative natural prices while, on the other, rates of return are uni-

form, yet changing, multiple and none is easily linked with money in the absence of a *produced* money commodity. As Garegnani has pointed out to me, it would not be *logically* impossible to extend the classical analysis in this direction but gains for the theory of accumulation are not obvious. Sraffa did not use "dated" prices in *Production of Commodities* in spite of several allusions to structural change (inventions, expansion of production to inferior lands etc.), and he attributed the origin of different own-rates of interest to changing "market prices" in his review of Hayek.

It is well known that a logical critique of the neoclassical theory of distribution using aggregate measures of capital has been attempted on the basis of the classical theory. I do not want to discuss here how other new versions of the neoclassical theory of distribution are affected by it. But the generality of the classical theory is enhanced by the fact that it is compatible with several theories of distribution which may each be more or less pertinent in different historical circumstances. To start from a given real wage and to regard profits as a residual was a first approach taken at a time when real wages had shown a remarkable stability over centuries in spite of short-term fluctuations. Others have argued that imperfect competition determines a degree of monopoly so that we should first explain the share of profits in national income which then determines the rate of profit. The Cambridge school has stressed the role of demand that, given different savings propensities of capitalists and workers, effects a redistribution of incomes between classes such that—within limits—a higher ratio of investment to output can be financed because savings rise with a rising share of profits. In this view, accumulation is so dynamic that prices can be kept above costs for considerable stretches of time in spite of the forces of competition. Firms are thus able to finance their expansion out of profits, that are high because the demand—largely generated by the firms themselves—is booming. Sraffa has suggested that the rate of profit is determined from outside the system of physical reproduction by the level of the monetary rates of interest. A very simple

version of this theory would hold in a situation of slow accumulation where competitive firms are indebted to the banking sector so that their profits must cover interest costs while competition keeps prices down and real wages absorb what is left of the net product.

The price theory under discussion can therefore not be subsumed as a special model under the neoclassical theory. It is general in that it is compatible with a number of alternative theories of distribution and output which have not yet all been formalized with the same degree of rigor and depend in their applicability on historical circum-

stances, but which, taken together, are related as variants of the classical theory and cover a wide range of empirical phenomena.

#### REFERENCES

- Hahn, Frank, "The Neo-Ricardians," *Cambridge Journal of Economics*, December 1982, 6, 353-74.
- Sraffa, Piero, "Dr. Hayek on Money and Capital," *Economic Journal*, March 1932, 42, 42-53.
- , *Production of Commodities by Means of Commodities*, Cambridge: Cambridge University Press, 1960.



# Joan Robinson's Critique of Equilibrium: An Appraisal

By E. ROY WEINTRAUB\*

Joan Robinson devoted much of her energy to critiquing equilibrium analysis, or what she was accustomed to calling Walrasian economics. In her various writings she returned again and again to the idea that Walrasian economics necessarily misdirected analysis in macroeconomics, capital theory, distribution theory, and value theory. Many of her writings made mention of Walrasian equilibrium, usually in the context of the wrong way to look at a particular problem. Indeed, for Robinson, the Walrasian theory came to stand for the intellectual opposition, as it were, for the "others" who had got it wrong, where the "it" was anything from the measurement of capital to the role of expectations in long run investment decisions.

My task today is a simple one. First I shall identify the central elements of Robinson's attack on Walrasian economics and then sketch what modern authors now understand to be the main points of the "neo-Walrasian research program." Finally, I shall compare Robinson's understanding of, and critique of, that program with the program itself. From this, I shall argue that Robinson's analysis flowed from her flawed understanding of the nature and role of equilibrium in the neo-Walrasian program.

## I. Robinson on Equilibrium

In their excellent survey article, Harvey Gram and Vivian Walsh note that "the theorist who would fairly appraise Robinson's work must immediately come to terms with the correct interpretation of her many attacks upon general equilibrium theory and comparative static analysis. She means by general equilibrium theory the static and intertemporal models of supply and demand equilibrium in the post-Walrasian tradition..." (1983, p. 519).

In "History and Equilibrium," Robinson noted that Frank Hahn had identified the competitive equilibrium as a set of prices such that, were they to obtain, no agent would have any incentive to modify any decision. She then stated:

This entails that everyone knows exactly and in full detail what consequences would follow any action that he may take.... Equilibrium is described as the "end of an economic process"; the story is usually told of a group of individuals each with an "endowment" of ready-made goods or productive capacity of some specific kind. By trading and re trading in a market, each ends up with a selection of goods that he prefers to those that he started with. If we interpret this as an historical process, it implies that, in the period of past time leading to "today," equilibrium was not established.... A system of simultaneous equations need not specify any date nor does its solution involve history. But if any proposition drawn from it is applied to an economy inhabited by human beings, it immediately becomes self-contradictory. Human life does not exist outside history.... [1979, p. 49]

Robinson lodged two separate objections to the idea of a competitive equilibrium. First, she faulted the neo-Walrasian model for ignoring real life concerns that give legitimacy to the activity of doing economic analysis. And second, she blamed the neo-Walrasian idea of equilibrium for necessarily biasing the conclusions of economic arguments in favor of a particular political-moral order.

As an example of the first kind of attack, consider her "The Disintegration of Economics." She argued:

The theory of market equilibrium, with given "endowments" and given "tastes" for a specific list of commodities is

\*Professor of Economics, Duke University, Durham, NC, 27706

essentially static. It can accommodate accumulation and change only by making the assumption that buyers and sellers have "correct foresight" of the future course of prices. A world of correct foresight is not the world in which human beings live. [1979, p. 94]

To illustrate the second type of attack on equilibrium notions, consider the following passage, from the same paper:

The great claim of equilibrium theory was that it showed how scarce means are allocated between alternative uses in accordance with consumers' tastes. The existence of scarce means (materials, energy, cultivable land) has recently come to the fore in public discussion, while consumers' tastes run to large cars, overheated rooms, and an excessive consumption of meat. The central doctrine of orthodox economics is the defence of the freedom of anyone who has money to spend, to spend it as he likes. [p. 92]

These illustrations of Robinson's criticisms of the neo-Walrasian idea of equilibrium merely sample the "identification she makes between neo-neoclassical economics—the object of many of her attacks—and post-Walrasian general equilibrium theory..." (Gram and Walsh, p. 519). It is not necessary here to examine other cases of Robinson's annoyance with the idea of equilibrium as it is used in neo-Walrasian analysis. It is necessary, however, to be quite clear about the object of her derision, for if her understanding of neo-Walrasian analysis is faulty, her attacks on it may be without merit.

## II. The Neo-Walrasian Research Program

As I have shown elsewhere (1985) it is helpful to think of general equilibrium analysis as a scientific research program in the sense in which the term was first used by the philosopher, Imre Lakatos. (See Mark Blaug, 1980.)

A research program in the sense of Lakatos is a constellation of theories linked by certain elements. That is, all the theories are

based on a certain set of presuppositions shared by all who work in the program. The organizing center of the program is called the hard core, which contains propositions taken as given by adherents to the program. The heuristics of the program suggest the ways that theories can be developed from the hard core. The heuristics are used to develop theories in the protective belts of the program. The core is never subjected to testing; rather it is the induced theories that are the appropriate subjects for corroboration and falsification.

As with any Lakatosian research program, the neo-Walrasian program is characterized by its hard core, heuristics, and protective belts. Without asserting that the following characterization is definitive, I have argued that the program is organized around the following propositions: HC1 *economic agents have preferences over outcomes*; HC2 *agents individually optimize subject to constraints*; HC3 *agent choice is manifest in interrelated markets*; HC4 *agents have full relevant knowledge*; HC5 *observable outcomes are co-ordinated, and must be discussed with reference to equilibrium states*.

By definition, the hard-core propositions are taken to be true and irrefutable by those who adhere to the program. "Taken to be true" means that the hard-core functions like axioms for a geometry, maintained for the duration of study of that geometry.

The positive and negative heuristics of the program link the hard core with the basic work of theory construction and development. The standard process of empirical conjectures and refutations is associated with the protective belts of the program. The heuristics of the neo-Walrasian program consist of propositions like: PH1 *construct theories in which economic agents optimize*; PH2 *construct theories that make predictions about changes in equilibrium states*; NH1 *do not create theories based on irrational behavior*; NH2 *do not create theories in which equilibrium has no meaning*; NH3 *do not test hard-core propositions*; and so forth.

The heuristics link the hard core to the theories in the belts that are the fit subject for testing and refinement. I submit that theories like human capital theory, "rational

expectations macroeconomics," the theory of black-white earnings differentials, and the theory of optimum tariffs, are theories in the protective belts of the neo-Walrasian research program. These theories are all based on the hard core, and the refinements and developments of these theories are guided by the heuristics of the program. Notice that we do not test the hard core propositions, as we do not test (with the idea of confirmation or rejection) the inverse square law in Newtonian mechanics, or the fixity of the earth's position with respect to the sun in the Ptolemaic program in astronomy—these are hard-core propositions that are maintained by workers in the program. The issue of whether, and when, a program (defined by the hard core) is "rejected" is an issue of the relative progress or degeneration of the program *with respect to a competing research program*. There is no evidence to suggest that economists abandon degenerating programs in the absence of a progressive alternative. We do not, in the face of falsified theories in the belt of a program, abandon that program until there is an alternative program with theories that are themselves corroborated.

The perspective of the neo-Walrasian program allows us to see that the notion of a "Walrasian" equilibrium is indeed fundamental to the enterprise that Robinson called "neo-neoclassical" economics (which roughly corresponds to my term "the neo-Walrasian program"). Using the same hard core and heuristics, neo-Walrasian economists create models of investment behavior, study the effect of open market operations on the level of real output and employment, and explain wage differentials between rural and urban workers in developing economies. All these economists, in their studies, use models which *interpret* the terms "agent," "optimize," "constraint," "knowledge," "market," and "equilibrium" in ways appropriate to the particular model. A coherent analysis will require a demonstration that an equilibrium exists for the model thus constructed. Predictions, and empirical corroborations or falsifications of those predictions, will be induced by the equilibrium concept that is consistent with the specification of the model.

### III. Robinson's Critique: An Appraisal

Returning to Robinson's critique of the neo-Walrasian idea of equilibrium, it is absolutely clear that her first set of objections to the notion of equilibrium is based on a profound misunderstanding of the concept of equilibrium and its function in analysis. The preceding section has shown that coordinated outcomes, equilibria, are imperfectly interpreted terms of the hard core of the neo-Walrasian program. Thus it is an organizing feature of any theory in the program that there must be a well-defined idea of equilibrium present; that equilibrium notion is to be intrinsic to any model or theory that exists in the protective belts.

It is a categorical mistake to complain that the Arrow-Debreu-McKenzie (model specific) competitive equilibrium is unrealistic or that it assumes knowledge of the consequences that would follow any action that an agent would take. Such corroborative tests are appropriate for theoretical constructs associated with the protective belts, but are not appropriately directed to the terms and structures of the hard core. All that Robinson's criticisms demonstrate is that she is not used to engaging in neo-Walrasians analysis, something that people perhaps knew already. To repeat, her criticisms of the idea of equilibrium itself, which focus on its presumed lack of realism, only identify Robinson as one who works in a different program.

Many of Robinson's attempts to convict neo-Walrasians of logical error are flawed in this way. The equilibrium notion is not testable, it is simply present as an organizing feature of the theories in the protective belts of the program. A monetary theoretic model's "perfect foresight equilibrium" can neither be corroborated nor falsified. Yet it is not meaningless to work with such equilibria if they lead to propositions that are potentially falsifiable.

From this perspective it becomes clear that Robinson adhered to an alternative program, perhaps one that would be termed the post-Keynesian program, and her criticisms of the neo-Walrasian program must be read as rhetorical set-pieces designed to gather new ad-

herents to her preferred program; that is, her arguments are for the most part exhortations, not exercises in theoretical or empirical logic.

Robinson also argued that the neo-Walrasian program biases the kinds of questions that economists can ask, and answer. In so arguing, for example, she identified economists who believe in the value of neo-Walrasian equilibrium notions with people whose "tastes run to large cars, overheated rooms and an excessive consumption of meat." This line of argument is suggestive of a moralist, not an analyst. Curiously, preaching *is* one of the few ways that one *can* attack a progressive research program like the neo-Walrasian program. Recent work by Arjo Klammer (1984) and Don McCloskey (1983) holds that such rhetoric is less disreputable in science than a rationalist's construction of the growth of knowledge would suggest.

Put another way, the choice which a scientist makes between two competing research programs will be based, using a rational reconstruction of the growth of knowledge, on the relative progressivity of the programs themselves. But such reasoned judgments are, in practice, not often found. The programs may compete for adherents with arguments that transcend reason. Moral suasion may be appropriate to coax a potential convert to the cause, the emergent program. Whether the convert finds the new religion congenial may later depend on the actual progressivity of the program as tested by the theoretical and empirical progress it demonstrates, but in the first instance, converts may be won by emotion, not reason.

#### IV. Conclusion

As a matter of logic, Joan Robinson's criticism of the nature and role of equilibrium in the neo-Walrasian program cannot withstand close scrutiny. Assailing the logic of hard-core propositions, as she did, is an exercise based on misunderstanding. As an attempt to gather adherents to a competing research program, Robinson's moralizing is explicable. The ultimate issue of course is the relative progressivity of the neo-Walrasian and post-Keynesian research programs. What is required to judge the issue is a comparative appraisal of those two competing research programs. It is a bit troubling that no such appraisal has as yet been attempted.

#### REFERENCES

- Blaug, Mark, *The Methodology of Economics*, New York: Cambridge University Press, 1980.
- Gram, Harvey and Walsh, Vivian, "Joan Robinson's Economics in Retrospect," *Journal of Economic Literature*, June 1983, 21, 518-50.
- Klammer, Arjo, *Conversations With Economists*, Brighton: Wheatsheaf Press, 1984.
- McCloskey, Don, "The Rhetoric of Economics," *Journal of Economic Literature*, June 1983, 21, 481-517.
- Robinson, Joan, *Collected Works*, Vol. 5, Oxford: Basil Blackwell, 1979.
- Weintraub, E. Roy, *General Equilibrium Analysis: Studies in Appraisal*, New York: Cambridge University Press, 1985.

## OPEN AND SEALED-BID AUCTIONS†

### Auction Theory with Private Values

By ERIC S. MASKIN AND JOHN G. RILEY\*

For many centuries, auctions have been a common form of selling procedure. Although auction methods vary across country and product, the two most frequently observed are the open, ascending bid (or English) auction and the sealed-bid auction. Recent theoretical research has led to a theory of equilibrium bidding in these two auctions and a wide range of alternatives as well. As a result it has been possible to compare the revenue extracted by the seller under different auction methods and even to characterize the revenue-maximizing auction.

The Revenue Equivalence Theorem (see for example, William Vickrey, 1961, Roger Myerson, 1981, and Riley and William Samuelson, 1981) asserts that when each bidder's reservation price for a unit of an indivisible good is an independent draw from the same distribution, and bidders are risk neutral, the sealed-bid auction generates the same expected revenue as the open auction. Much recent research has involved weakening each of the main hypotheses—risk neutrality, identically distributed values, and independence of values—in turn. We shall illustrate some of the principal conclusions of this work by considering the properties of open and sealed-bid auctions in a model of two bidders whose reservation prices can assume only two values, and by comparing these auctions to the “optimal” or revenue-maximizing auction.

†*Discussant:* William F. Samuelson, School of Management, Boston University.

\*Departments of Economics; Harvard University, Cambridge, MA 02138, and University of California, Los Angeles, CA 90024, respectively. We are indebted to William Samuelson for helpful suggestions. We thank the Sloan Foundation and the NSF for financial support.

#### I. Revenue Equivalence

Imagine that the reservation price of bidder  $i$  ( $i=1,2$ ) can assume the values  $v_H$  (with probability  $p$ ) and  $v_L$  (with probability  $1-p$ ), where  $v_H > v_L \geq 0$ . Bidders' values are private information and independently distributed. Bidders are risk neutral, that is, they maximize the expression

$$(1) \quad (\text{probability of winning})v \\ - \text{expected payment.}$$

We suppose that an open auction proceeds by the auctioneer's continuously raising the asking price. The auction concludes when one of the bidders drops out. The remaining bidder is the winner and pays the dropout price (if both bidders drop out simultaneously, a coin is flipped to determine the winner). Given these rules, one can easily confirm that a bidder's unique (perfect) equilibrium strategy is to drop out when the asking price reaches his reservation price. (There are other “nonperfect” equilibria, see our 1983a paper). Thus the expected payoff of a  $v_L$  bidder (a bidder whose reservation price is  $v_L$ ) is zero, and his probability of winning is  $\frac{1}{2}(1-p)$ . The expected payoff of a  $v_H$  bidder, by contrast, is his surplus if the other bidder is “low” (since then the asking price only reaches  $v_L$  rather than  $v_H$ ) times the probability of that event, that is,  $(1-p)(v_H - v_L)$ . Since a  $v_H$  bidder wins when the other bidder has a low value and wins half the time when the other bidder has a high value, his probability of winning is  $\frac{1}{2}p + (1-p)$ .

In the sealed-bid auction, bidders submit bids simultaneously. The higher bidder is the winner (ties again are resolved by coin flips) and he pays his bid. Consider a symmetric



equilibrium. Because the distribution of values is discrete, the equilibrium will involve mixed strategies. Notice first that a  $v_L$  bidder (one whose reservation price is  $v_L$ ) will never bid more than  $v_L$  because, if he did, the maximum of such bids (if bidders use mixed strategies that randomize over a variety of alternative bids) would win the auction with positive probability, inducing a negative expected payoff. Let  $\underline{b}_L$  be the infimum of all bids submitted.

Suppose first that  $\underline{b}_L < v_L$ . Then bidders bid below  $v_L$  with positive probability and so a  $v_L$  bidder's expected payoff is positive. Suppose, furthermore, that bidder 1 bids  $\underline{b}_L$  with positive probability. Then bidder 2's chances of winning increase discontinuously if he bids just more than  $\underline{b}_L$  while his payment if he wins scarcely rises, thereby raising his expected payoff. But this is a violation of symmetry. On the other hand, if  $\underline{b}_L$  is not bid with positive probability, then bids near  $\underline{b}_L$  have almost no chance of winning, contradicting the positive expected payoff.<sup>1</sup>

Next let  $\underline{b}_H$  be the infimum of bids made by a  $v_H$  bidder. If  $\underline{b}_H > v_L$ , then a bid strictly between  $\underline{b}_H$  and  $v_L$  has the same chance of winning as  $\underline{b}_H$ , and so is preferable. Thus  $\underline{b}_H = v_L$ , and a  $v_H$  bidder's expected payoff must be  $(v_H - v_L)(1 - p)$ . In equilibrium, any bid  $b$  made as part of a mixed strategy must generate the same expected payoff. Therefore if  $F(b)$  is the cumulative distribution function of a  $v_H$  bidder's bid, it satisfies

$$(2) \quad [pF(b) + 1 - p](v_H - b) = (1 - p)(v_H - v_L).$$

By symmetry, a  $v_H$  bidder's expected probability of winning is  $\frac{1}{2}p + (1 - p)$ , whereas that of a  $v_L$  bidder is  $\frac{1}{2}(1 - p)$ . Because a given type of bidder's probability of winning and expected payoff are the same in the open and sealed-bid auctions, formula (1) implies that his expected payment is the same in the two auctions. We have established the Reve-

nue Equivalence Theorem for our model. Indeed, we obtain the same expected revenue from any other auction in which the high bidder wins, the expected payoff of a  $v_L$  bidder is zero, and the expected payoff of a  $v_H$  bidder is  $(1 - p)(v_H - v_L)$ .

It is of some interest to compare the open and sealed-bid auctions with a revenue-maximizing auction (see Myerson and Riley-Samuelson). Suppose that bidders were offered the choice between bidding  $v_L$  or  $b_H = (\frac{1}{2}v_H + \frac{1}{2}(1 - p)v_L)/(\frac{1}{2}p + 1 - p)$ , with, as always, the high bidder winning. Because  $b_H$  is greater than  $v_L$ , a  $v_L$  bidder will bid  $v_L$ . Since  $(\frac{1}{2}p + 1 - p)(v_H - b_H) = \frac{1}{2}(1 - p)(v_H - v_L)$ , a  $v_H$  bidder is indifferent between bidding  $b_H$  and  $v_L$ , and so might as well choose the former. Since a  $v_H$  bidder bidding  $b_H$  has the same probability of winning as in an open or sealed-bid auction ( $\frac{1}{2}p + (1 - p)$ ), but has a lower expected payoff,  $(\frac{1}{2}(1 - p)(v_H - v_L))$  rather than  $(1 - p)(v_H - v_L)$ , his expected payment must be higher. Thus, this alternative auction generates higher expected revenue. Indeed, it is optimal if  $v_L > pv_H$ . (If  $v_L < pv_H$  it is optimal to set a reserve price at  $v_H$ , thereby rejecting all lower bids.) In either case, the optimal auction differs from the open and sealed-bid auctions by prohibiting bidders from making certain bids. This conclusion generalizes to more complicated models, including those with a continuum of possible reservation prices.

## II. Risk Aversion

Let us modify the model of Section I only by supposing that bidders are risk averse. Let  $u$  be a strictly concave von Neumann-Morgenstern utility function, normalized so that  $u(0) = 0$ . A  $v$  bidder's payoff if he wins and pays  $t$  is  $u(v - t)$ ; his payoff if he loses and pays  $t$  is  $u(-t)$ .

Risk aversion does not alter the bidders' behavior in the open auction; it is still optimal for a bidder to drop out exactly when his reservation price is reached. Hence expected revenue is as before. In the sealed-bid auction,  $v_L$  bidders continue to bid  $v_L$ , and if  $F_R$  is the cumulative distribution function

<sup>1</sup>Our argument here presumes that the equilibrium in the sealed-bid auction is symmetric. One can show (see our 1983a paper) that there is no asymmetric equilibrium.

of a  $v_H$  bidder's bid, it satisfies the analogue of condition (2):

$$(3) \quad u(v_H - b)[(1 - p) + pF_R(b)] \\ = u(v_H - v_L)(1 - p).$$

The strict concavity of  $u$  implies that  $u(v_H - v_L)/u(v_H - b) < (v_H - v_L)/(v_H - b)$  for  $v_L < b < v_H$ . Hence, (2) and (3) imply that  $F_R(b) \leq F(b)$  with strict inequality for bids greater than  $v_L$  but less than the maximum. That is,  $F_R$  stochastically dominates  $F$ , and so the expected bid by a  $v_H$  bidder is higher with risk aversion than without. We conclude that, with risk aversion, a sealed-bid auction generates greater expected revenue than an open auction (see Gerard Butters, 1975, and Charles Holt, 1980). Intuitively, increasing a bidder's risk aversion heightens his fear of losing and so, in a sealed-bid auction, induces him to bid higher. Viewed alternatively, a sealed-bid auction, unlike an open auction, insures a winning bidder against fluctuations in the amount he has to pay, and a risk-averse bidder is willing to pay a premium—in the form of a higher bid—for this insurance.

By requiring payments even of losing bidders, an optimal auction (see our 1984 article, and Steven Matthews, 1983) can exploit the fact that a risk-averse bidder's marginal utility of income depends on whether he wins or loses. Let  $\pi_i$  be the probability of winning and  $b_i$  and  $a_i$  the payments by a winning and losing bidder, respectively, of type  $i$  ( $i = L, H$ ). An optimal auction chooses  $\pi_i$ ,  $b_i$ , and  $a_i$  to maximize

$$(4) \quad p(\pi_H b_H + (1 - \pi_H) a_H) \\ + (1 - p)(\pi_L b_L + (1 - \pi_L) a_L),$$

subject to

$$(5) \quad \pi_H u(v_H - b_H) + (1 - \pi_H) u(-a_H) \\ \geq \pi_L u(v_H - b_L) + (1 - \pi_L) u(-a_L)$$

$$(6) \quad \pi_L u(v_L - b_L) + (1 - \pi_L) u(-a_L) \geq 0$$

$$(7) \quad \frac{1}{2}p + (1 - p) \leq \pi_H$$

$$(8) \quad \frac{1}{2} \geq p\pi_H + (1 - p)\pi_L$$

$$(9) \quad \pi_H \geq 0 \quad \text{and} \quad \pi_L \geq 0.$$

Constraint (5), a self-selection constraint, ensures that a  $v_H$  bidder is at least as well off making a high as a low bid. We have omitted the analogous self-selection constraint for a  $v_L$  bidder since, as we shall see, it is satisfied automatically. Constraint (6) guarantees a  $v_L$  bidder a nonnegative expected payoff from participating. (Given (5), a  $v_H$  bidder's payoff will also be nonnegative.) Condition (7) says that a  $v_H$  bidder can win with at most probability 1 if the other bidder has a low reservation price and, given the symmetry of the model, with at most probability  $\frac{1}{2}$  if the other bidder's reservation price is high. Constraint (8) requires simply that each bidder's probability of winning, *unconditional* on his reservation price, not exceed  $\frac{1}{2}$ .

Letting  $\alpha$  and  $\beta$  be the Lagrange multipliers for (5) and (6), respectively, we obtain the first-order conditions

$$(10) \quad p\pi_H - \alpha\pi_H u'(v_H - b_H) = 0$$

$$p(1 - \pi_H) - \alpha(1 - \pi_H) u'(-a_H) = 0$$

$$(11) \quad (1 - p)\pi_L + \alpha\pi_L u'(v_H - b_L)$$

$$- \beta\pi_L u'(v_L - b_L) = 0$$

$$(1 - p)(1 - \pi_L) + \alpha(1 - \pi_L) u'(-a_L)$$

$$- \beta(1 - \pi_L) u'(-a_L) = 0.$$

From (10) we find that  $v_H - b_H = -a_H$ , that is, a high bidder is perfectly insured; he receives a monetary transfer  $-a_H$  ( $> 0$ ), as compensation if he loses. From (11) and the fact that  $u'(v_H - b_L) < u'(v_L - b_L)$ ,

$$(12) \quad (\beta - \alpha) u'(-a_L)$$

$$= 1 - p > (\beta - \alpha) u'(v_L - b_L).$$

Thus a  $v_L$  bidder is better off winning than losing ( $v_L - b_L > -a_L$ ). Moreover, since (from (12)) (6) is binding, he must actually pay a penalty if he loses ( $a_L > 0$ ), which we

can interpret  $a$  as an entry fee. Because (5) is binding and  $v_H - b_L > -a_L$ , we have  $v_H - b_H < v_H - b_L$ , that is, a  $v_H$  winner pays more than a  $v_L$  winner. If (8) is binding, as it will be if  $p$  is small enough, we can solve for  $\pi_L$  and rewrite (4) as  $p\pi_H(b_H - a_H - b_L + a_L) + pa_H + (\frac{1}{2} - p)a_L$ . From the above argument,  $b_H - a_H - b_L + a_L > v_H - v_L > 0$ . Hence, constraint (7) is binding:  $\pi_H = \frac{1}{2}p + (1 - p)$ .

We conclude that an optimal auction with risk-averse bidders resembles that for risk-neutral bidders. Bidders are offered the choice between two prices  $b_H$  and  $b_L$  (if, as before,  $p$  is not too high), and the high bid wins. However, if a bidder loses with a bid of  $b_H$ , he is compensated for losing, whereas if he loses with a bid of  $b_L$ , he is penalized. Intuitively, introducing a penalty heightens a risk-averse bidder's fear of losing and therefore increases the revenue that can be extracted from a  $v_H$  bidder. Of course, this penalty, by increasing risk, reduces the payment that a  $v_L$  bidder makes. But the penalty has no effect to the first-order, since, with no penalty, a  $v_L$  bidder is perfectly insured.

It remains only to show that the solution to the program of maximizing (4) subject to (5)–(9) satisfies

$$(13) \quad \pi_L u(v_L - b_L) + (1 - \pi_L)u(-a_L) \\ \geq \pi_H u(v_L - b_H) + (1 - \pi_H)u(-a_H),$$

the self-selection constraint for  $v_L$  bidders. But (13) follows immediately from the fact that (5) holds with equality and  $\pi_H u'(v - b_H) > \pi_L u'(v - b_L)$  (since  $\pi_H > \pi_L$  and  $b_H > b_L$ ) for all  $v$ .

### III. Asymmetry

Let us revert to risk neutrality but now drop the assumption that valuations are identically distributed. Specifically, assume that bidder 1's reservation price is distributed as in Section I, but that bidder 2's reservation price is either  $w_H$  or  $w_L$  with probabilities  $q$  and  $1 - q$ , respectively. Continue to suppose that the two bidders' distributions are independent. For convenience, let us suppose that  $v_L = w_L = 0$ . Then the

expected revenue generated by the open auction is

$$(14) \quad pq \min\{v_H, w_H\}.$$

We wish to compare the difference in revenues,  $\Delta$ , between the sealed-bid and open auctions.<sup>2</sup> To do this we shall consider two polar cases of asymmetry: (i) both bidders have the same probability of being high but have different high values, that is,  $p = q$  and  $v_H \neq w_H$ , and (ii) both bidders have the same high values but different probabilities, that is,  $v_H = w_H$  and  $p \neq q$ .

It is not difficult to see that in case (i),  $\Delta$  is positive. We know from Section I that when  $v_H = w_H$ ,  $\Delta$  is zero. Now imagine raising  $w_H$  above  $v_H$ . This does not affect revenue from the open auction since there is no change in the distribution of the second highest reservation value. However, with a higher  $w_H$ , the optimal response in the sealed-bid auction by bidder 2 (when  $v = w_H$ ) to bidder 1's equilibrium strategy is a higher bid. Bidder 2's higher bid, in turn, induces bidder 1 to bid higher than before (for details, see our 1983b paper). Hence, revenue from the sealed-bid auction rises, and  $\Delta$  becomes positive.

In case (ii), expected revenue in the open auction is  $pqv_H$ . In the sealed-bid auction, the equilibrium cumulative distribution functions,  $F_1$  and  $F_2$ , of the bids of bidders 1 and 2, when their reservation prices are  $v_H$ , satisfy the analogue of (2):

$$(15) \quad (1 - q + qF_2(b))(v_H - b) \\ = (1 - q + qF_2(0))v_H;$$

$$(16) \quad (1 - p + pF_1(b))(v_H - b) \\ = (1 - p + pF_1(0))v_H.$$

<sup>2</sup>As our model is formulated, an equilibrium in the sealed-bid auction may not exist. The nonexistence problem, however, is an artifact of our allowing literally a continuum of possible bids. In fact, we can restore existence even with a continuum by allowing the possibility of positive but infinitesimal bids, which we implicitly assume in our analysis.

Notice that right-hand sides of (15) and (16) allow for the possibility that a  $v_H$  bidder will bid zero (actually, infinitesimally more than zero) with positive probability. This will be the case if  $p \neq q$  since both bidders must make the same maximum bid,<sup>3</sup>  $\bar{b}_H$ , when their reservation price equals  $v_H$ , and (15) and (16) can be satisfied for  $b = \bar{b}_H$  only if one of  $F_1(0)$  and  $F_2(0)$  is nonzero. For example, if  $p > q$ , then (15) and (16) imply that

$$\bar{b}_H = qv_H = pv_H(1 - F_1(0)),$$

and so  $F_1(0) = 1 - q/p$ . Integrating (15), we obtain  $qv_H$  as the expected payment by bidder 1 if his reservation price is  $v_H$ , where  $z = \int F_2(b) dF_1(b)$ . Similarly, from (16), the expected payment by bidder 2 is  $(p(1 - z) + q - p)v_H$ . Hence total expected revenue is  $q^2v_H$ , which is less than the open auction revenue,  $pqv_H$ . Therefore, for case (ii),  $\Delta$  is negative.

Roughly speaking, the sealed-bid auction generates more revenue than the open auction when bidders have distributions with the same shape (but different supports), whereas the open auction dominates when, across bidders, distributions have different shapes but approximately the same support.

#### IV. Correlation

Let us return to the model of Section I, except now assume that reservation prices are correlated across bidders. Specifically, let  $r_{ij}$  ( $i, j \in \{L, H\}$ ) be the joint probability that bidder 1's value is  $v_i$  and that bidder 2's value is  $v_j$ . Correlation implies that

$$(17) \quad r_{HH}r_{LL} - r_{HL}r_{LH} \neq 0.$$

As usual, behavior in the open auction remains the same, and so expected revenue is

$$(18) \quad r_{HH}v_H + (1 - r_{HH})v_L.$$

Making the obvious modifications in the

analysis of Section I, we conclude that expected revenue for the sealed-bid auction is also (18). This equivalence between the two auctions does not generalize to distributions with more than two point supports because, in general, with correlation, a higher reservation price does not imply a higher bid for the sealed-bid auction (although it does for the open auction).<sup>4</sup> Any condition sufficient to guarantee that bids are monotonic in reservation prices, however, ensures equivalence. One such condition is that the reservation prices be affiliated (see Paul Milgrom and Robert Weber, 1982).

When (17) holds, an optimal auction extracts all surplus from bidders (see Jacques Cr  mer and Richard McLean, 1985). To see this, let  $c_{ij}$  ( $i, j \in \{L, H\}$ ) be the payment that bidder 1 makes when his  $v = v_i$  and bidder 2's  $v = v_j$ . To extract all surplus, the  $c_{ij}$ s must satisfy

$$(19) \quad \frac{1}{2}r_{LL}v_L - r_{LH}c_{LH} - r_{HH}c_{HH} = 0$$

$$(20) \quad (\frac{1}{2}r_{LH} + r_{LL})v_L - r_{LH}c_{HH} - r_{LL}c_{HL} < 0$$

$$(21) \quad (\frac{1}{2}r_{HH} + r_{HL})v_H - r_{HH}c_{HH} - r_{HL}c_{HL} = 0$$

$$(22) \quad \frac{1}{2}r_{HL}v_L - r_{HH}c_{LH} - r_{HL}c_{LL} < 0.$$

Equations (19) and (21) require the surplus of  $v_L$  and  $v_H$  bidders, respectively, to be zero. Inequality (20) ensures that a  $v_L$  bidder is not better off bidding as a  $v_H$  bidder, and (22) imposes the corresponding constraint on a  $v_H$  bidder. But from (17), we can solve for  $c_{ij}$ s that satisfy (19)–(22).

<sup>4</sup>Suppose, for example, that  $v$  can take on three possible values:  $v_H > v_M > v_L$ . Assume that if  $v = v_H$  for one bidder, then it is very likely that  $v = v_L$  for the other bidder. Assume further that if  $v = v_M$  for one bidder, then the other bidder in all likelihood has the same reservation price. In this case, a  $v_M$  bidder will bid higher on average than a  $v_L$  bidder in the sealed-bid auction. Furthermore, the sealed-bid auction, at least for some parameter values, generates strictly more revenue than does the open auction.

<sup>3</sup>If, say, bidder 1's maximum bid were greater than that of bidder 2, bidder 1 could lower his bid without reducing his probability of winning.

### V. Concluding Remarks

We have discussed three major hypotheses of the Revenue Equivalence Theorem, but there remain two more implicit in our formulation. One is the assumption that only a single item is sold. If buyers have downward-sloping demand curves and there are multiple units for sale, the Revenue Equivalence Theorem again fails. Extrapolating from some simple examples, we conjecture that open bidding will tend to dominate sealed bidding in this environment.

The second assumption is that a bidder's reservation price does not affect the reservation price of any other bidder. This is the "private values" hypothesis: the assumption that reservation prices are a matter of taste rather than a reflection of information about the intrinsic value of the good. In the latter case, the "common values" model, the open auction tends to produce higher revenue than the sealed-bid auction when our other hypotheses are maintained (see Milgrom and Weber).

### REFERENCES

- Butters, G. R., "Equilibrium Price Distributions and the Economics of Information," unpublished doctoral dissertation, University of Chicago, 1975.
- Cr  mer, J. and McLean, R. P., "Optimal Selling Strategies Under Uncertainty for a Discriminating Monopolist When Demands Are Interdependent," *Econometrica*, 1985, 53, forthcoming.
- Holt, C., "Competitive Bidding for Contracts under Alternative Auction Procedures," *Journal of Political Economy*, June 1980, 88, 433-45.
- Maskin, E. S. and Riley, J. G., (1983a) "Uniqueness of Equilibrium in Open and Sealed Bid Auctions," mimeo., April 1983.
- \_\_\_\_\_ and \_\_\_\_\_, (1983b) "Auctions With Asymmetric Beliefs," Discussion Paper No 254, University of California-Los Angeles, June 1983.
- \_\_\_\_\_ and \_\_\_\_\_, "Optimal Auctions With Risk Averse Buyers," *Econometrica*, November 1984, 52, 1473-518.
- Matthews, S. A., "Selling to Risk Averse Buyers With Unobservable Tastes," *Journal of Economic Theory*, August 1983, 30, 370-400.
- Milgrom, P. and Weber, R. J., "A Theory of Auctions and Competitive Bidding," *Econometrica*, September 1982, 50, 1089-122.
- Myerson, R., "Optimal Auction Design," *Mathematics of Operations Research*, 1981, 6, 58-73.
- Riley, J. G. and Samuelson, W. F., "Optimal Auctions," *American Economic Review*, June 1981, 71, 381-92.
- Vickrey, W., "Counterspeculation, Auctions and Competitive Sealed Tenders," *Journal of Finance*, March 1961, 16, 8-37.

# Empirical Testing of Auction Theory

By ROBERT G. HANSEN\*

Given the state of affairs in auction theory—there is at least one model, for instance, to support any position one would care to take concerning the revenue of sealed-bid vs. open auctions—it should not come as a surprise that a fair amount of empirical work in auctions is underway. This paper reports the results of some recently completed research. I first discuss papers in which the predictions being tested derive directly from the pure theory of auctions, and then papers in which the predictions arise out of an application of auction theory to a related institution.

## I. Tests of the Pure Theory

I consider here two implications for which we now have a fair amount of empirical evidence: the prediction that sealed-bid and open auctions yield equal revenue for a seller; and the prediction that individual bids in a sealed-bid auction decrease with the number of bidders. These predictions are not, of course, robust across all auction models. It is therefore somewhat brave to speak of testing “auction theory” as if there were one theory giving us one unambiguous prediction. Although much of the empirical work is presented as classical hypothesis testing, it is then probably better to think about that work as informal Bayesian learning that is only guided by the structure common to all auction models (for example, Nash equilibrium for a given number of expected utility-maximizing bidders).

## II. Revenue Results of Open vs. Sealed-Bid Auctions

Implications concerning revenue appear to be the most easily testable predictions of

auction theory; all that is required is a suitable data set. To my knowledge, the best data set available covers U.S. Forest Service sales of contracts for harvesting timber in the Pacific Northwest during 1977. Because of a change in federal law, sealed bids accounted for about half of the several hundred sales in that year; open auctions, the usual procedure, made up the rest.

There has been considerable analysis of this data both within the Forest Service and under contract for the Forest Service. At the best, this work consists of regressions using individual sales as observations, the high bid as dependent variable, and a list of independent variables that includes a dummy variable for auction method, the number of bidders, and several variables controlling for the quality of the sale. The general finding is that sealed-bid auctions yield significantly greater revenue than open auctions (for instance, Walter Mead et al., 1981, report roughly a 10 percent higher price for sealed-bid auctions).

My own work on this data set (1984a) corrects for two problems contained in the earlier findings: first, there was some selection bias in the way the Forest Service chose auction methods; and second, auction theory implies a test for revenue equivalence that is different from simply using one dummy variable to account for auction method.

With respect to the first problem, it became obvious upon a little researching that a correct model for these auctions would be a simultaneous equations model, with one equation—a probit model—describing the auction-choice process, and a second equation determining the high bid conditional on auction method. The fact that certain unobservable variables (for example, firms' timber inventories) affected both auction method and high bid leads to nonzero covariance for the error terms of these equations; estimation techniques will therefore have to be something other than *OLS*.

The second problem with early work concerns how auction method enters the high-bid

\*Amos Tuck School of Business Administration, Dartmouth College, Hanover NH 03755. I thank my dissertation committee, especially John Riley, for academic support, and the Sloan Foundation and Tuck Associates for financial support.

equation. Resolution of this problem required a simple reexamination of some basic assumptions in auction theory. Specifically, theorists always compare revenue for a fixed and known number of auction participants. Upon reflection, this implies—with a positive reserve price—that the actual number of *bidders* is not the variable that the theoreticians have in mind. Indeed, if one looks at data for one-bidder sales, one sees that sealed-bid auctions clearly dominate open auctions. That this says nothing about revenue equivalence follows from the observation that the number of bidders is not an accurate measure of the number of traders as envisioned by theorists. (Actually, this evidence on one-bidder sales supports an implicit assumption of auction theory—that participants in a sealed-bid auction do not learn who has a reservation value exceeding the reserve price before the submission of bids.)

The bottom line is that an accurate high-bid model should account for the number of actual bidders and the number of potential bidders. Furthermore, it is likely that the effect of actual bidders on the high bid will differ across auction methods (thus, calculating expected revenue differences entails taking expected values over both the regression error term and over the number of actual bidders, and classical hypothesis testing for revenue equivalence requires testing that several coefficients are jointly zero).

As it turns out, only the first concern—that of neglecting selection bias—seems to affect the empirical results. Whether or not the number of potential bidders is controlled for, *OLS* estimation suggests an expected high-bid difference (sealed less open) of about \$15 (per thousand board feet) with a standard error of around \$3. I used two alternative methods to correct for selection bias, a two-stage method as suggested by Maddala and full-information maximum likelihood. These methods imply an expected revenue difference of between \$1 and \$6, with a standard error of \$5. For comparison purposes, the overall average high bid is about \$130. Also, the joint hypotheses implied by revenue equivalence cannot be rejected at the 95 percent level.

I also analyzed the data in more detail to check for recognizable subsets where any

difference between the auctions would be particularly pronounced. There are some theoretical results on asymmetry of beliefs and on collusion that suggest certain areas as potential recognizable subsets. The empirical procedure here was to take a variable indicating possible asymmetry or ease of collusion (for instance, the concentration ratio for timber purchase by region was used to measure ease of collusion), and then look at the average residuals from the high-bid equation for varying levels of this proxy variable. If, for instance, open auctions make cartels more stable, then open auctions should underperform sealed-bid auctions, especially where the ease of collusion is low. This will show up as a pattern in residuals.

As it turned out, no recognizable subsets could be found even though several proxy variables were used. My conclusion is that anyone with strong revenue equivalence priors should not be shaken.

### III. Competition and Bid Levels

There has been considerable work showing that high bids, no matter what the auction method, increase with the number of bidders. My work of the previous section is just one example. A more interesting problem is whether individual bids also increase with the number of bidders. The “winner’s curse” suggests they might not: with more bidders, there is more negative information associated with winning the auction (this is only within a common-value context).

The most sophisticated analysis of this question is the work of G. W. Gilley and G. V. Karels (1981). They use data from U.S. offshore oil lease auctions and an estimation procedure that takes into account a problem similar to one I encountered for timber auctions: with a positive reserve price, some traders will not submit bids even though they place positive expected value on the tract. Gilley and Karels show that using data only on observed bids biases the estimates because of usual truncated-error considerations. In a model using individual sealed bids as the dependent variable, correcting for this problem via Heckman’s procedure changes the coefficient on “number of bidders” to significantly negative from significantly posi-

tive under *OLS* estimation. Oil firms seem to understand the winner's curse rather well.

#### IV. Tests of the Applied Theory

I briefly discuss some empirical results based on applications of auction theory to the means-of-payment in auctions, and to the effectiveness and strategies of cartels.

Taking the means-of-payment issue first, several results can be derived concerning the use of stock and cash-stock bids in the market for corporate control. First, if we make independent-preference assumptions for how potential acquiring firms value a target firm, it turns out that stock bidding will yield the target more revenue than cash bidding. Interesting as this may be, it seems difficult to test. If we make additional assumptions concerning private knowledge, however, more easily testable predictions fall out.

For example, if the target has private knowledge on its value, it is simple to show that stock and cash-stock offers alleviate the resulting adverse-selection problem encountered by acquirers. Importantly, however, this holds true only when the target firm is not too small relative to the acquirers—if too small, the stock payment doesn't have the desirable contingent-payment feature. Of course, the reverse holds true when it is the acquiring firms that have private knowledge on their values: then it will be cash which alleviates adverse selection and facilitates trade.

A directly testable prediction is then that stock trades should be seen with greater frequency for transactions where the target is not small relative to the acquirer. Although formal tests have yet to be done, preliminary evidence supports this prediction. Further tests could be done if it were possible to identify transactions by the likelihood of target-side or acquirer-side asymmetry. In general, testing models based on private knowledge in such a direct fashion is sure to be difficult; it would seem generally preferable to devise indirect tests such as the size-based test above. This notwithstanding, there is one piece of evidence that can be interpreted as supporting the model: target firms with high market-value to book-value

ratios are most likely to be acquired in stock transactions (see Willard Carleton et al., 1983). This constitutes supportive evidence if we accept the argument that high market-to-book ratios imply the existence of significant intangible assets, the value of which is inherently highly uncertain and susceptible to private knowledge.

Turning to cartel behavior, the presumption has been around for some time that sealed-bid auctions will make life tougher for a bidder's cartel than will open auctions. Marc Robinson (1983) has formalized this and shown that a sealed-bid auction with reserve price is a seller's best defense against cartels. However, this work is again empirical only to the extent that it gives one reason for the widespread use of sealed bid auctions with a reserve price.

Recent work by Jonathan Feinstein, Michael Block, and Frederick Nold (1985) goes one step further. They show how a bidder's cartel can, in an intertemporal auctions context, extract additional monopoly rents from the seller by bidding in such a way as to affect the seller's beliefs about expected future costs and thereby induce him to sell more contracts now. For instance, by keeping the variance of bids low, the cartel can lead the seller into believing that today's bid is a good estimate of what next period's will be. The implications of the model are supported by evidence from confirmed cartel behavior in auctions of North Carolina highway contracts. Specifically, not only are bids on average low when cartel behavior is present, but bids also vary less.

#### V. Outlook

The outlook for additional statistical work in auctions appears good. New theoretical results are coming out for sequential auctions; these may be testable using data from, for instance, Michigan's open, sequential oil lease auctions. Also, for offshore oil lease auctions and Forest Service timber auctions (sealed-bid), data on the whole distribution of bids can be obtained. Following the lead of Feinstein et al., further theoretical results on bid distributions could be formulated and tested. And last, there is the chance that



somebody will again experiment with different auctions.

#### REFERENCES

- Carleton et al., Willard T., "An Empirical Analysis of the Role of the Medium of Exchange in Mergers," *Journal of Finance*, June 1983, 38, 813-27.
- Feinstein, Jonathan S., Block, Michael K. and Nold, Frederick C., "Asymmetric Information and Collusive Behavior in Auction Markets," *American Economic Review*, June 1985, forthcoming.
- Gilley, G. W., and Karels, G. V., "The Competitive Effect in Bonus Bidding: New Evidence," *Bell Journal of Economics*, Autumn 1981, 12, 637-49.
- Hansen, Robert G., "Auctions with Contingent Payments," *American Economic Review*, June 1985, forthcoming.
- \_\_\_\_\_, (1984a) "Empirical Testing of Auction Theory," Working Paper No. 161, Tuck School, 1984.
- \_\_\_\_\_, (1984b) "Informational Asymmetry and the Means-of-Payment in Auctions," Working Paper No. 162, Tuck School, 1984.
- Mead, Walter J., Schniepp, Mark and Watson, Richard B., "The Effectiveness of Competition and Appraisals in the Auction Markets for National Forest Timber in the Pacific Northwest," U.S. Forest Service Contract No. 53-3187-1-43, 1981.
- Robinson, Marc S., "Oil Lease Auctions: Reconciling Economic Theory With Practice," Department of Economics Working Paper No. 292, University of California-Los Angeles, May 1983.

# Experimental Development of Sealed-Bid Auction Theory; Calibrating Controls for Risk Aversion

By JAMES C. COX, VERNON L. SMITH, AND JAMES M. WALKER\*

We offer a brief survey of bidding theory in high price auctions, of experimental studies of behavior in such auctions, and of the interplay between the design and results of the experiments and efforts to further develop the theory. Two new series of experiments are reported. The first applies a convex transformation of payoffs in an attempt to induce a lowering of subject bids "as if" the bidders had become less risk averse. The second applies a method for inducing any prespecified utility function (for risky choices) on an individual. We use it to induce "as if" risk-neutral behavior. Both series use baseline control to "calibrate" the hypothesized effect of the procedures on "risk-averse" behavior.

## I. Bidding Theory and its Development under Testing

Early experimental papers testing William Vickrey's (1961) noncooperative equilibrium model of bidding behavior for risk-neutral agents in single unit auctions report the robust result that subjects tend to bid significantly higher than the predictions of the model when the number of bidders is  $N \geq 4$ , but not when  $N = 3$ . (For citations to our experimental-theoretical work, see the references in our 1984 article.) The results for  $N \geq 4$  are consistent with extensions of the Vickrey model which postulate that agents all have the same concave utility for monetary surplus (for example, Charles Holt, 1980). However, these extensions also imply

that all bidders use the same equilibrium bid function:  $b_i(v_i) = b(v_i) \geq b_n(v_i)$ , for all  $i$ , where  $v_i$  is the value of the auctioned item (known only) to  $i$  and  $b_n(v_i) = (N-1)v_i/N$  is the Vickrey risk-neutral bid function when each  $v_i$  is drawn independently from the constant density on  $[0, \bar{v}]$ . We have tested the null hypothesis that the bids submitted by the  $N$  bidders in each experimental group can be regarded as  $N$  samples from the same population. It is rejected in 13 of 23 experimental groups. A straightforward conclusion is that an appropriate extension of the model should be based on the assumption of heterogeneous risk-averse bidders. We have articulated such a model for single unit auctions and extended it to multiple unit discriminative auctions. Experimental tests of the multiple unit model strongly support the interpretation that bidders bid as if they were heterogeneous and risk averse (we reject the hypothesis of homogeneous agents in 24 of 28 experimental groups).

This constant relative risk-averse (CRR) model assumes that (a) each agent  $i$  chooses  $b_i$  to maximize  $EU(b_i) = (v_i - b_i)^{r_i} G_i(b_i)$ , where  $G_i(b_i)$  is the probability that  $b_i$  is the highest of  $N$  bids; (b) agent expectations are rational,  $G_i(b_i) = G(b_i)$ ; (c) each  $v_i$  in any auction is drawn independently from the constant density on  $[0, \bar{v}]$ ; (d) the  $N$  agents are drawn from a population with some distribution  $\Phi(r_i)$  on the characteristic  $r_i \in (0, 1]$ . For single unit auctions, these assumptions imply the inverse equilibrium bid function

$$(1) \quad v_i = (N-1+r_i)b_i/(N-1),$$

for all  $b_i \in [0, \bar{b}]$ ,

where  $\bar{b} = (N-1)\bar{v}/N$  is the maximum bid that would be made by a risk-neutral agent. (The solution for  $b_i > \bar{b}$  has no closed form.) Hence, if any two of  $N$  bidders ( $i, j$ ) have

\*Cox and Smith: Professors, University of Arizona, Tucson, AZ 85721; Walker: Assistant Professor, Indiana University, Bloomington, IN 47401. Financial support by the National Science Foundation (grants SES-8205983, SES-8404915, and SES-8310096) is gratefully acknowledged.

distinct *CRR*A parameters ( $r_i, r_j$ ), a prediction of the model is that in a sequence of auctions  $t = 1, 2, \dots, T$ , the observed bids will identify a distinct homogeneous linear bid function for each bidder whose slope will reveal each bidder's *CRR*A utility parameter.

We have conducted and reported two direct tests of the above *CRR*A model and its multiple unit generalization. The first test applies the following property of *CRR*A utility: a scalar change,  $\alpha$ , in payoffs has no effect on expected *CRR*A utility-maximizing decisions. This is seen in (a) above if we express expected utility in the form,  $EU(b_i) = [\alpha(v_i - b_i)]^{\alpha} G(b_i)$ , where  $\alpha(v_i - b_i)$  is the outcome in U.S. currency for the winning bidder. Since  $\alpha$  affects only the scale of utility it has no effect on the bid function (1). We report paired comparison experiments in which  $\alpha = \$1$  in the control experiments and  $\alpha = \$3$  in the paired treatment experiments. There is no significant difference in the outcomes between paired experiments. Similarly, if  $\bar{v}$  is increased, say tripled, this model predicts the same scalar increase in  $v_i$ ,  $b_i$  and  $v_i - b_i$ ; that is, in (a) we can write expected utility in the form  $EU(\mu_i) = [\bar{v}(v_i - \mu_i)]^{\alpha} G(\mu_i)$ , where  $\mu_i = b_i/\bar{v}$ ,  $v_i = v_i/\bar{v}$ . Hence a scalar change in the  $v_i$  has no effect on normalized bids. We have reported comparison experiments (multiple units) in which  $\bar{v}$  (and each  $v_i$ ) is tripled. The effect is to triple average bids, and the conclusions based on  $\alpha = \$1$  are not altered when  $\alpha = \$3$ .

Since *CRR*A utility is the only utility function with these scalar invariance properties, these experiments provide important independent support for the theory beyond the earlier *ex post* analysis showing that observed bids are consistent with the assumption that agents are risk averse and heterogeneous. The new tests for scalar effects were motivated a priori by the theory.

Two well-known "logical" objections to all *CRR*A utility models are a recurring part of the conventional wisdom connected with expected utility theory (*EUT*), although these objections are devoid of observational support: (A) "*CRR*A utility is unacceptable as it implies that absolute risk aversion grows without bound as  $v - b$  approaches zero";

(B) "The *CRR*A bidding model admits of a tractable solution only if initial wealth is (or can be normalized on) zero."

The a priorist objection (A) asserts that any *EUT* model can be "tested" by examining its absolute risk-averse implications, and that behavior near some boundary (zero) is a crucial test of any hypothesis. This is like arguing (without resorting to observation) that the inverse square law of attraction is falsified, since the force of attraction goes to infinity as the distance between masses approaches zero! For the vast majority of subjects, when  $v$  is near zero, the ratio of bid to value is similar to that for large values of  $v$ ; that is, one observes no peculiarity in bidding behavior near zero, which is predicted by equation (1) based on *CRR*A utility. A small minority of subjects bid either zero or their value, at low values of  $v$ . This "throw away" bid phenomenon can be interpreted as the result of payoffs being so low that it is not worth the trouble of a "serious" bid. Since it is not clear what is "optimal" when payoffs are at *epsilon* levels, other theories such as random or erratic behavior should not be discounted, just as in particle theory (which is disciplined by data) other theories take over in the small.

Concerning objection (B), we have been quite explicit from the beginning in referring to  $v_i - b_i$  as the monetary *income* from an auction. This is because we accept the findings of a vast literature going back at least to Markowitz, and corroborated by Mosteller and Nogee, Davidson, Suppes and Siegel, Edwards, Kahneman and Tversky, Binswanger, and others (see Mark Machina, 1982, for numerous references). Generally this literature supports the relative invariance of risk-taking decision behavior with initial wealth (the "Markowitz hypothesis" of a horizontally shifting utility of wealth). Also, this literature does not find support for constant absolute risk aversion (*CARA*) (Machina, p. 285). *The prize to which EUT applies* (wealth, income etc.) *is a hypothesis separate from the axioms of EUT which do not define that prize.*

Various extensions of the original Vickrey model and of Holt's identical bidders risk-averse model are contained in the literature.

Paul Milgrom and Robert Weber (1982) consider the effect of information. However, given the robust experimental result that bidders bid as if they are heterogeneous and risk averse, these extensions (for first price and Dutch auctions) are based on behavioral assumptions already shown to be inconsistent with the data. Eric Maskin and John Riley (1984) offer a potentially fruitful extension based on the assumption of a one-parameter utility function  $u(-b_i, \theta_i)$ , where  $b_i$  is  $i$ 's bid. If  $\theta_i = v_i$ , we have the case of identical bidders with differing private values, but with the bidding commodity medium distinct from the commodity item being auctioned. If  $\theta_i = r_i$  (any risk parameter), we have heterogeneous risk-averse bidders, but, implicitly, all must place the same value on the auctioned item. Thus the minimum generalization requires a utility function of the form  $u(v_i - b_i; r_i)$  to capture both taste and risk attitude diversity. Cox and Smith (1984) develop an equilibrium bidding model for a utility function of the form  $u(\theta_1 - b, \theta_{-1})$ , where  $(\theta_1, \theta_{-1})$  is an  $M$  vector of characteristics.

## II. Models of Control for the Effect of Risk Aversion

We interpret the observation that subjects bid in excess of the predictions of the Vickrey model as due to heterogeneous risk-averse agents, and have used this *interpretation* to develop an improved model. Subsequently this model was found to be consistent with the scalar invariance tests described above. Now we ask whether direct methods might be applied to examine this risk-averse interpretation of the data. Other interpretations are possible. We might assume in place of (a) that agents choose bids to maximize  $EU(b_i) = (v_i - b_i)G_i(b_i)$ , and instead of (b), that  $G_i(b_i) = [G(b_i)]^{1/r_i}$ , where  $1/r_i$  is now a characteristic of bidder  $i$  that transforms the objective probability of winning,  $G(b_i)$ , into a subjective probability of winning,  $[G(b_i)]^{1/r_i}$ . This subjective expected value (SEV) model is prominent in psychology (see Machina, pp. 290-91). It abandons Muthian rational expectations, but the resulting model yields a bid function identical with (1), and the two theories are observa-

tionally equivalent on the basis of all experimental tests to date. The methodological point is that the parameter  $r_i$  is not observable; it is a construct based on an interpretation of what is driving behavior, and other interpretations are potentially admissible. We have adopted the heterogeneous risk-averse interpretation because it is an integral part of the traditional *EUT*, while the alternative is thought to be "*ad hoc*." This does not mean that *EUT* is "true," but that it appears that there is not yet a sufficient basis for the scientific community to abandon *EUT*.

We propose two payoff manipulation models which, based on *EUT*, should have a determinate effect as interpreted in terms of risk aversion. If these models are "correct," and our interpretation that subjects are risk averse is correct, the new data should be consistent with these predictions.

*Model I.* In a first price auction, if subject  $i$  wins, suppose that instead of paying  $(v_i - b_i)$  dollars to  $i$  we pay  $a(v_i - b_i)^2$  dollars,  $a > 0$ . In the *CRRRA* model it is seen that the problem now is to maximize  $EU(b_i) = [a(v_i - b_i)^2]^{r_i} G(b_i)$  and equation (1) becomes  $v_i = (N - 1 + 2r_i)b_i / (N - 1)$ , for all  $b_i \in [0, \hat{b}]$ , where  $\hat{b} = (N - 1)\bar{v} / (N + 1)$ . Thus if an individual's personal measure of *CRRRA* is  $1 - r_i$ , under the payoff transformation of Model I, that individual will behave "as if" the *CRRRA* measure had changed to  $1 - 2r_i$ . This equation provides strong quantitative predictions of the effect of the transformation. A weaker qualitative prediction is that the individual will bid less under the transformation.

*Model II.* Instead of paying  $(v_i - b_i)$  dollars to the high bidder, suppose we pay the winner  $(v_i - b_i)$  unit lottery tickets. The individual then participates in a lottery in which he/she receives  $x_1$  dollars in U.S. currency with probability  $(v_i - b_i)/\bar{v}$  and  $x_2$  dollars ( $x_1 > x_2$ ) with probability  $1 - (v_i - b_i)/\bar{v}$ . Suppose further that the  $N - 1$  low bidders in the auction all receive  $x_2$  dollars. Since the probability of  $x_1$  is linearly increasing in  $(v_i - b_i)$ , if *EUT* applies to individual behavior this procedure will cause the individual to bid "as if" risk neutral (Alvin Roth and Michael Malouf, 1979; Joyce Berg et al., 1984). To see this, note that bidder  $i$ 's deci-

sion problem is to

$$\begin{aligned}
 (2) \quad & \max_{b_i} \left[ G_i(b_i) \left\{ u_i(x_1) \left( \frac{v_i - b_i}{\bar{v}} \right) \right. \right. \\
 & \left. \left. + u_i(x_2) \left( 1 - \frac{v_i - b_i}{\bar{v}} \right) \right\} + (1 - G_i(b_i)) u_i(x_2) \right] \\
 & = \left[ \frac{u_i(x_1) - u_i(x_2)}{\bar{v}} \right] \\
 & \times \max_{b_i} [(v_i - b_i) G_i(b_i)] + u_i(x_2),
 \end{aligned}$$

which is formally equivalent to Vickrey, giving  $v_i = Nb_i/(N-1)$ , if  $G_i(b_i) = G(b_i)$ .

Berg et al. have generalized this procedure to induce any prespecified preferences, which they proceed to test experimentally using *CARA* and constant absolute risk-prefering (*CARP*) preferences. They provide a qualitative test by soliciting responses to a choice between two bets, *A* and *B*, with the property that the induced *CARA* function implies that *A* is preferred to *B* while the *CARP* function predicts that *B* is preferred to *A*. They report that significantly more than half (88.3 percent) of the choices correspond to the predictions. They also elicit minimum selling prices for bets from the two groups and compare these with the calculated certainty equivalents of the bets. The observed prices reported by the subjects are then compared with those predicted to provide a quantitative test of their model. For both groups they reject the null hypothesis of no relationship between the *average* observed prices and those predicted. However, the average prices from the risk-averse group tended to be systematically biased above the predicted certainty equivalent. Also, the variance of the observed prices was high.

This is encouraging in that it provides evidence favoring the gross predictive implications of inducing known preferences on subjects. The procedure is potentially important in enabling one to (a) control for risk aversion where other aspects of behavior are the primary focus of the investigation (Roth and Malouf), or (b) induce known risk pref-

erences in a market whose behavior is hypothesized to be driven by risk aversion.

To our knowledge this promising procedure has not been test-calibrated in a market context; that is, used to induce particular preferences in a market which yields predictions *interpretable* in preference terms. High price auctions allow one to do this based on the Vickrey risk-neutral special case.

### III. Experiments and Results: Model I, Quadratic Transformation

Twelve subjects participated in three sessions each consisting of 4 bidders. Each session consisted of a baseline sequence (*EIB*) of 20 auctions (12 in session 1) in which each subject was paid one cent for each *PLATO* experimental cent earned,

$$(3) \quad (\text{cash cents}) = (\text{PLATO cents}),$$

followed by a transformation sequence (*EIT*) of 20 auctions in which for each auction cash earnings were calculated using

$$(4) \quad (\text{cash cents}) = 0.02 (\text{PLATO cents})^2.$$

In the *PLATO* instructions for *EIT*, tables and graphs are used to inform the subjects of the payoff implications of this transformation. After the first three sessions were completed, four of the subjects were recruited for a fourth retest session consisting of 20 transformation auctions.

Our initial approach to comparing bidding behavior in *EIT* and *EIB* was twofold. First, we ask whether the mean normalized bid of a subject differs in a transformation experiment from that in a baseline experiment. Since the value realizations from the uniform distribution will differ in the two experiments, if *i* bids  $b_i^*$  when the realized value is  $v_i^*$ , we normalize the bid by subtracting the risk-neutral Vickrey bid,  $b_n(v_i^*)$ . Thus, for each *i* we compute the difference  $D_i = b_i^* - b_n(v_i^*)$  for each auction, giving a set of differences  $\{D_i^B\}$  in *EIB* and a set  $\{D_i^T\}$  in *EIT*. The means  $\bar{D}_i^B$  were positive for all subjects, indicating that all were risk averse in the baseline sequence. Also,  $\bar{D}_i^T > 0$  for all

$i$ , as each subject continues to exhibit risk aversion under the transformation. Furthermore,  $\bar{D}_i^B - \bar{D}_i^T$  is positive for eight of the twelve subjects indicating, as predicted, a shift toward less risk-averse behavior with the transformation. In the retest, session 4, all four subjects bid lower in *EIT* than in *EIB* (one subject had bid higher in the earlier session).

However, this apparently good support for Model I could not withstand deeper examination. The payoffs in (3) and (4) imply that a bid for a profit of less than 50 yields a lower return under the transformation than in the baseline, and vice versa for a bid with potential profit in excess of 50. Aware of this, in advance of the experiments we had conjectured that a "satisficer" might bid relatively lower (higher) in *EIT* at profit levels below (above) 50, as a means of maintaining *EIB* performance in *EIT*. A closer examination of individual bids revealed that this effect was strong, contrary to the predictions of Model I.

#### IV. Experiments and Results: Model II, Lottery Payoffs

Twelve new subjects participated in three sessions, each consisting of 4 bidders and two parts. The first part was a sequence of 20 baseline experiments (*E2B*) with each subject paid one cent in cash for each cent earned in the experiment. The second part consisted of a sequence of 20 auctions (*E2L*) in which subjects in effect earned one lottery ticket for each cent won in an auction. Eight of the twelve subjects were then recruited for two retest sessions, 4 and 5, consisting of 20 auctions with the lottery payoff. The lottery operated as follows: A box containing 1000 tickets, numbered consecutively, was displayed to the subjects. The high bidder in each auction was assigned ticket numbers in an amount equal to the bidders experimental profit in cents. Thus, if the winning bidder had a value of \$8 and bid \$6 when  $\bar{v} = \$10$ , she might be assigned the lottery numbers 1–200. If the ticket she then drew was in the range 1–200, she received a cash payoff of \$7.50. Otherwise, she received \$0.25. All losing bidders received \$0.25 in cash.

Ten of the twelve subjects bid *significantly* "as if" risk averse in *E2B*, but, contrary to Model II, only one of these subjects was induced to bid as if risk neutral in *E2L*. It appears that Model II has one chance in ten of making a correct strong form prediction. This result was not changed for any subject in the retest sessions 4 and 5. A weak form prediction of Model II is that the difference between baseline and lottery mean bids,  $\bar{D}_i^B - \bar{D}_i^L$ , is positive indicating a shift toward risk neutrality. Only six of twelve subjects were consistent with this prediction.

#### V. Conclusions of the Calibration Experiments

Model I, applying a quadratic payoff transformation, predicts a doubling of the "as if" *CRR*A parameter  $r_i$ , or, more weakly, a shift in the direction of lower bids (less risk-averse bidding). The experimental results belie this prediction. A close examination of individual bidding suggests that subjects bid less only when the profit potential is below the 50 cent "break-even" level. Above this 50 cent potential profit level, subjects tend to bid relatively higher. This can be interpreted as a type of "satisficing" behavior in which subjects attempt to do at least as well under the quadratic transformation as in the baseline experiments. Does this test invalidate the *CRR*A model of bidding? No; literally, it questions the *conjunction* of the *CRR*A model with the transformation of Model I. Since the *CRR*A model has performed well in previous tests, Model I should be the immediate focus of deeper examination. In particular, the results suggest the need for a change in design that would eliminate the break-even 50 cent profit defined by the intersection of the baseline and quadratic payoff functions. The predicted result is that the hypothesized satisficing effect will be eliminated. Of course, the theory asserts that behavior should not be affected by this artifact, but one would like to know if the theory does better when the artifact is removed. After all, these are not calculating agents, and it may not be difficult to introduce perceptual distortions that alter equilibrium behavior.

Model II predicts risk-neutral bidding for any subject showing risk-averse bidding in a baseline experiment. Since nine of ten subjects bid significantly above the risk-neutral bid function under both lottery and monetary payoffs, these results do not support the predictions of Model II. Given the generally supportive results of earlier direct tests of the bidding model, the predictive failure of Model II can be interpreted as providing (indirect) evidence against the compound lottery axiom of *EUT* that is essential in Model II. Furthermore, these results may have implications for other research programs that must postulate the behavioral validity of the lottery procedure as a conditional in experimental tests of models that require risk attitude of agents to be controlled.

#### REFERENCES

- Berg et al., Joyce E., "Controlling Preferences for Lotteries on Units of Experimental Exchange," Working Paper 1983-5, Graduate School of Management, University of Minnesota, February 1984.
- Cox, James C. and Smith, Vernon L., "Equilibrium Bidding Theory When Some Bidders May Be Risk-Preferring," University of Arizona, 1983, rev. August 1984.
- \_\_\_\_\_, \_\_\_\_\_, and Walker, James M., "Theory and Behavior of Multiple Unit Discriminative Auctions," *Journal of Finance*, September 1984, 39, 983-1010.
- Holt, Charles A., "Competitive Bidding for Contracts Under Alternative Auction Procedures," *Journal of Political Economy*, June 1980, 88, 433-45.
- Machina, Mark J., "Expected Utility Analysis Without the Independence Axiom," *Econometrica*, March 1982, 50, 277-323.
- Maskin, Eric and Riley, John, "Optimal Auctions With Risk Averse Buyers," *Econometrica*, November 1984, 52, 1473-518.
- Milgrom, Paul R., and Weber, Robert J., "A Theory of Auctions and Competitive Bidding," *Econometrica*, September 1982, 50, 1089-122.
- Roth, Alvin E. and Malouf, Michael W. K., "Game-Theoretic Models and the Role of Information in Bargaining," *Psychological Review*, November 1979, 86, 574-94.
- Vickrey, William, "Counterspeculation, Auctions, and Competitive Sealed Tenders," *Journal of Finance*, March 1961, 16, 8-27.

## Modes of Thought in Economics and Biology

By PAUL A. SAMUELSON\*

Sociobiology, particularly in connection with its obsessive interest in Hamiltonian altruism, has veered closer towards economics. Reactions within economics against highfalutin borrowings of the methodology of mathematical physics led Alfred Marshall and a host of later writers to hanker for a "biological" approach to political economy. A dispassionate audit of what resulted from this urge would have to report a disappointing paucity up to now of fruitful finds or insights.

But independently of this anti-physicism strain in economics there developed within economics an "evolutionary approach." It is not social Darwinism that I am referring to (although an insidious temptation in that direction is sometimes involved). Rather, several writers explored the notion that—what survives under competition, whether in the jungle or the marketplace, may resemble what is achieved in a maximum problem's solution—even though no participants in the struggle may have any perception and awareness that they as individuals are maximizing anything or awareness that the group ends up doing so. Some names associated with this notion are Armen Alchian (1950), Milton Friedman (1953), and Sidney Winter and Richard Nelson (1982). My *Foundations* (1947) discussions of the stability of surviving forms, borrowed explicitly from the physiologist philosopher L. J. Henderson (1917), is somewhat in this same vein.

There is really no trace of anti-physicism in such an approach. Indeed, the case where

soap bubbles unconsciously form the shapes that maximize elaborate integrals in the calculus of variations is very much like the case where leaves and branches of a tree, competing for access to the sunlight, achieve maximal shapes.

When Edward Wilson's manifesto for sociology first came out (1975), I predicted that it would meet a resonant response in Chicago-school economics. Everything has proceeded on schedule to validate my forecast. A hard-boiled economist who shares his income with his child can now cease to feel guilty about this superficial violation of the doctrine of "no free lunch." Hamiltonian inclusive fitness shows that this is just a selfish gene's way of making more of itself—no less respectable a process than that in which I cannily outsell or outproduce my neighbor.

### I. Population as Part of Economics

Population analysis is at the intersection of social science and biology. Once upon a time, throughout the heyday of classical economics, demography belonged to political economy. The supply of labor was one of the important endogenous variables in the systems of Smith, Malthus, Mill, and Marx. I have expressed in many places how superficial and empty the concept of a subsistence wage has always been. Although economies in the pre-industrial revolution era differed widely in their per capita levels of real income, writers blithely embraced the paradigm of a horizontal *SS* supply curve for labor, as if the cost of production and reproduction of labor was an identifiable exogenous parameter. There must be something very tempting about this vacuous notion, since so many different scholars did succumb

<sup>†</sup>*Discussant:* Gordon Tullock, Public Choice Center, Virginia Polytechnic Institute and State University.

\*Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139. I owe thanks to the Sloan Foundation for partial support.



to it. And, by the age of Mill and Marx—indeed, even in Ricardo's time and Malthus's later editions—the level of the subsistence wage rate was freed from physiological dimensions and became a dummy variable shot through with conventional standards of life hankered after by (some) workers. At this stage a meaningful false theory had been replaced by a meaningless nontheory that could never be vindicated or rejected by any pattern of historical facts.

## II. Paradigm Lost and Regained

One feature of neoclassical economics that distinguishes it from the classical version is the removal of population as a variable subject to economists' equilibrium analysis. Knut Wicksell, an autodidact in a backwater of northern Europe, was the last neoclassicist to preoccupy himself with demography. Marshall, Edwin Cannan, Frank Knight, and A. C. Pigou all were content to have the trend of population growth imported from the fields of sociology and technical demography. The pioneers of modern demography—Alfred J. Lotka, Robert R. Kuczynski, David Glass, Kingsley Davis, Ansley J. Coale, Nathan Keyfitz, and others—were not primarily economists and made little pretence to be. The classical subsistence wage theory is dead forever—dead as a theory of constancy of wage even as applied to the poor regions of Africa and Asia where gross reproduction rates are still enormously high and where population density perceptibly reduces per capita productivity. So the real wage rate is definitely not an exogenous variable for today's economics. Nevertheless, in recent years economists have begun to infiltrate the field of demography. This is part of the imperialist movement in which we economists try to apply our methodologies to everything—to the law, to the sociology of the family (courtship, marriage, divorce, cohabitation). I have in mind writings by such economists as Richard Easterlin, Paul Schultz, Gary Becker and a whole Chicago school, Ronald Lee, Harvey Leibenstein, and many others.

As might be expected, noneconomists have sometimes resented invasion of their turf.

Demographers are not as titillated as we are by letting the decision to have or not have a child be likened to the decision to buy or not buy a new car. Easterlin aside, the economists' models did not predict in advance the magnitude of the post-1939 baby boom; nor the post-1957 resumption of the trend toward small family sizes. Slutsky-like analysis of the income-and-family size relationship labors like a lion with concepts of *quality* and alternative earning power of spouses, only to produce the mouse that sometimes these days it is fashionable to have 1 or 2 children and sometimes to have 2 or 3. Weak and hard-to-reproduce evidences for extrasensory perception, people find boring. Similarly, a weak tendency for people to have more or to have less of their children during cyclical recessions does not excite noneconomists.

Economists are smart and hard-working people, now being produced in copious supply relative to the traditional problems feasible to make progress on. Like lemmings oppressed by the workings out of the law of diminishing returns, economists will continue to swarm into the area of demography. Hard work and intelligence will get themselves heard, even when somewhat hamstrung by unpromising initial hypotheses and methods. So, in the end, economists' demography will carve out an ecological niche for itself.

## III. Anecdotes and Anecdotes

A Kuhnian expert in the dynamics of scientific schools will observe how sociobiology extrudes into the traditional grounds of the social sciences. Francis Galton, Karl Pearson, and Ronald Fisher sought genetic explanations for differences in performance and class position. Although R. A. Fisher was a genius in genetics and mathematical statistics, his 1930 classic, *The Genetic Theory of Natural Selection*, seems naive in the degree to which he hypothecates a genetic basis for observed differences in behavior. No cogent evidence is given for his views and indeed it is not clear what evidence would be cogent.

We are beginning to identify today, by means of biochemistry and molecular biology, particular diseases with particular sin-

gularities in DNA data. But just how the courage observed in offspring is related to the genes for courage inherited from their ancestors is as yet a completely nonoperational question. The same has to be said of "altruism," "intelligence," "sexuality," and most of the attributes discussed in the final chapters of Wilson. These final chapters, I noticed at first reading, depend largely upon hypothetical anecdotes and case studies. Initially, I thought this a weakness of the material dealing with man and with higher primates, an understandable let down from the higher level of scientific rigor characterizing the earlier chapters dealing with the solid facts of animal and plant biology. But on further examination and reflection, I discovered that much of what goes under the name of natural history involves a similar marshaling of selective anecdotes. Chemistry and meteorology are not like that. And neither is economics. In economics we have data aplenty. Our subject is inexact but we do apply quantitative analysis (statistical and otherwise) to formulate and test our hypotheses. When you read about Volterra's struggle for existence between predators and prey, or Wilson's programs for a genetic-based ethics, you are not in the constrained world of fact and interpretative hypothesis but rather in the imaginative realm of speculative possibility.

To be sure, more and better observations will be made available by future scholars working in this vineyard. But the point to be made is how far away we now are from a satisfactory state of knowledge in these between-disciplines domains.

Rather than belabor these points about realism and relevance, I wish for the rest of the present endeavor to deal with purely deductive matters, to illustrate by some examples involving sex ratios how unlike (and alike) are the deductive paradigms common to economics (and perhaps sociology) and those of sociobiology and genetic demography.

#### IV. Different Moods

There are important differences between a typical genetic process and an economic or

sociological process. Genetic evolutionary selection, as when there is survivability advantage in a malarial environment for the genotype that is heterozygous for the sickle cell trait, involves no consciousness or overt purpose: any teleological aspect is an "as if" phenomenon. Ofttimes genetic change is glacially slow, involving hundreds of generations. So to speak, there is no "action at a distance" in demographic genetics: funeral by funeral, mating by mating, each allele must make its way. By contrast, in economics what one firm innovates to do now can in principle sweep the whole industry by quick imitation.

As a typical process in biology, consider a society that relies solely on the rhythm method for its birth control (a specification already shot through with sociology and convention). Given enough time the strategy will partially or wholly self-destruct—as those genetically disposed to have irregular menstrual cycles leave more of their offspring to be represented in subsequent populations. But the biological time scales involved run into hundreds of years, by which time there will have occurred scores of alternations of sociological ideologies.

The sex ratio is a useful topic for our comparison.

#### V. Balanced Sex Ratio

Human babies are almost as likely to be girls as boys: typically males are in small excess, 106 to 100; but females have lower mortality so that by adulthood the sex ratio is in virtual balance. Animal populations, whether monogamous or not, are observed often to have nearly balanced sex ratios at birth, with whatever unbalance that implies in maturity.

R. A. Fisher gives an ingenious argument why individual natural selection might be expected to evolve toward a balanced sex ratio at birth or conception. Here is my attempt to paraphrase his logic:

Suppose all genetic strains but yours tend to produce an excess of male over female births. If your strain tends to produce a more balanced sex ratio of

births, other things equal, each of you will be better represented with grandchildren than each of the other crowd — this for the reason that, for producing the grandchildren generation, the totality of *all* sons have exactly equal representation with the totality of *all* daughters (even when the respective totals are unequal!), with the implication that each daughter brings you more in relative grandchildren representation than each son does. So, in each generation, the unbalance of the sex ratio at birth tends to be reduced. Only a sex ratio at birth in balance is immune to evolutionary drift.

To see how mathematical geneticists put this heuristic argument on a rigorous analytical basis, refer to I. Eshel (1975), Joel Yellin and myself (1977), and the Gibbs Lecture of Samuel Karlin (1984). To illustrate the logic it suffices for me here to sketch the rock-bottom simplest case.

#### VI. Darwin's Mindless Invisible Hand

Adult male and females,  $[M, F]$  mate to produce this period's births,  $B$ . The fractions  $[g, 1-g]$  determine the breakdown of newborns between males and females. The mortality fractions  $[p_m, p_f]$  specify how many of the newborn males and females survive to be adults in the next period. With mortality specified, our model is complete once we specify the mating-fertility function determining the output of births out of the inputs of adult males and females. I write this as

$$(1) \quad B(t) = M(t)F(t)b[M(t), F(t)],$$

where  $Mfb[M, F]$  is much like the economists' production function, being monotone increasing (and perhaps concave). Examples would be  $cMF/(M+F)$ ,  $cMF/(\frac{1}{3}M + \frac{2}{3}F)$ ,  $cMF/M^{1/3}F^{1/2}$ .

Combine (1) with our mortality and sex-ratio assumptions

$$(2) \quad \begin{aligned} M(t) &= p_m [gB(t-1)], \\ F(t) &= p_f [(1-g)B(t-1)] \end{aligned}$$

to derive an autonomous first-order difference equation determining births and all demographic variables:

$$(3) \quad \begin{aligned} B(t) &= g(1-g)p_m p_f B(t-1)^2 \\ &\times b[p_m gB(t-1), p_f(1-g)B(t-1)] \\ &= \phi[B(t-1)], \quad t \geq 1. \end{aligned}$$

Thus, if  $Mfb[M, F]$  is first-degree homogeneous, we find Malthus-Lotka exponential growth:

$$(4) \quad \begin{aligned} B(t) &= B(0)R^t, \\ B(0) &= M(0)F(0)b[M(0), F(0)] \\ 1 \cong R &= g(1-g)p_m p_f [gp_m(1-g)p_f]. \end{aligned}$$

If both sexes played a symmetric role in the  $Mfb$  function, maximal group fitness would require that any superiority of females in mortality ought to be compensated by having an excess of males at birth  $g > \frac{1}{2}$ . The above Fisherine argument denies that the Invisible Hand of individual selection will lead to this group fitness state, leading instead to balanced sex ratio at birth ( $g = \frac{1}{2}$ ) whatever that implies for unbalance in the adult sex ratio!

Now introduce genetic selection. The old genetic strain that  $(M, F, B, g)$  above belonged to, I now write as  $(M_A, F_A, B_A, g_A)$ . A mutation creates also a new genetic strain, with  $(M_a, F_a, B_a, g_a)$ . (To keep the discussion simplest, I don't deal with diploid inheritance, which would require us to introduce  $[M_{AA}, M_{Aa}, M_{aa}, F_{AA}, \dots]$ , etc.) The axioms of the Fisherine syllogism are very strict:

1. The genotypes differ *only* in sex ratio:  $g_A \neq g_a$ .
2. The same survival fractions  $[p_m, p_f]$  apply to both genotypes.
3. Where fertility-mating is concerned, an individual of genotype  $A$  has *exactly* the same properties as any like-sexed, like-aged individual of genotype  $a$ . This implies random or nonassortative mating and that total

births,  $B = B_A + B_a$ , are determined as before by total  $(M, F)$  numbers:

$$(5) \quad B_A + B_a = (M_A + M_a)(F_A + F_a) \\ \times b[M_A + M_a, F_A + F_a].$$

For short, the last factor on the right will be written as  $b$ .

4. The last axiom specifies that offspring of parents both of the same genotype are of that genotype; half the offspring of parents of different genotype average out to be of the father's genotype, half of the mother's, and sampling irregularities are assumed ignorable in a large enough population.

The axioms translate in a straightforward way to give the following first-order difference equation system determining how the vector  $[B_A(t), B_a(t)] = [B_A, B_a]$  transforms itself into  $[B_A(t+1), B_a(t+1)] = [B'_A, B'_a]$ :

$$(6A) \quad B'_a = \{g_a(1-g_a)B_a^2 + \frac{1}{2}[g_A(1-g_a) \\ + g_a(1-g_A)]B_AB_a\} \{p_m p_f b\}.$$

$$(6a) \quad B'_A = \{g_A(1-g_A)B_A^2 + \frac{1}{2}[g_A(1-g_a) \\ + g_a(1-g_A)]B_AB_a\} \{p_m p_f b\};$$

Suppose we are interested only in ratios of genotypes,  $B_a/B_A = x(t)$ , and in the ratio of the sexes at birth, as measured by  $g(t)$  the properly weighted average of  $[g_A, g_a]$ . We can divide (6a) by (6A) and cancel out completely both any differences in the sexes' mortality fractions and any special features of the  $b[M, F]$  function that happen to obtain!

We then get the autonomous first-order difference equation in  $[x(t), x(t+1)]$ :

$$(7) \quad x(t+1) = x(t) \times \\ \frac{g_a(1-g_a)x(t) + \frac{1}{2}[g_A(1-g_a) + g_a(1-g_A)]}{g_A(1-g_A) + x(t)\frac{1}{2}[g_A(1-g_a) + g_a(1-g_A)]}$$

It can then be shown graphically or analytically that, for  $0 < x(0) < \infty$ , as  $t \rightarrow \infty$  the overall sex ratio  $g(t) \rightarrow \frac{1}{2}$  if  $(g_A, g_a)$  straddle

$\frac{1}{2}$ ; if they are both on one side of  $\frac{1}{2}$ ,  $g(t) \rightarrow$  the  $(g_A, g_a)$  nearest to  $\frac{1}{2}$ .

**Warning:** Balance of the sexes at birth would obtain in this model even if males lived longer than females,  $p_m > p_f$ , and even if females were more important for nurturing and if a small fraction of the available males would suffice for purposes of procreation. Fitness of the *group*, when it calls for balanced sex ratio of adults or such adult unbalance as implies a strong unbalance at birth, will be sacrificed to that maximizing *individual fitness*. Economists understand well that what's good for each may be bad for all sellers—as in Prisoner's Dilemma, or in perfect competition.<sup>1</sup>

## VII. Economic and Sociological Processes

New techniques are making it possible to know in advance what will be the sex of a baby. With knowledge may come power to control. Thus, if one learns by amniocentesis that an embryo is female, and if one prefers a male, an abortion might be decided on. If such customs become common, some far-reaching sociological changes could occur. And these would have economic ramifications.

Nor is all this new. Animal breeders and folklore have held that the probabilities of the different sexes can be affected by timing patterns within the menstrual cycle. Indeed, it is reported in *Too Many Women?* by M. Guttenberg and P. F. Secord (1983) that the

<sup>1</sup>See Fisher (pp. 142–43), W. D. Hamilton (1967), Wilson (pp. 316–17), R. L. Trivers and D. E. Willard (1973), for alternative models that alter conclusions about balanced sex ratios by bringing in considerations of “parental investment” and other interactions between genotype and fertility or mortality functions and parameters. Where an economist would say: “Parents will invest in one sex or the other up to the point of equality of *marginal advantage*,” biologists more often speak of equality of investments in the two sexes. Equality of derivatives and equality of ordinates are not always the same thing! Actually, selection would lead to two-thirds of newborns being male if the  $(M, F)$  symbols in (1)–(7), instead of meaning (males, females), had to be interpreted to be (pairs of males, single females)—as when each female uses up twice the “energy” or “mass” of one male.

Talmudic practices followed by nineteenth-century orthodox Jews of Eastern Europe resulted in a sex ratio as distorted as 146/100. Also, the mystery of the alleged excess of male births occurring in the World War I epoch may even be resolvable by prosaic factors having to do with scheduling of army leaves and discharges during that period.

Whatever the mixture of science fiction and fact characterizing present knowledge, it is not unlikely that better control over gender is just around the corner for reasons that have nothing to do with genetic change. How do economists model such a process? I shall deal only with an archetypal example.

Scarcity tends toward value. The more women there are, other things equal, the lower might be expected to be their relative wages and lifetime earnings. Feminists might therefore look with some approval upon a future trend toward more sons rather than daughters.

Working against this in the political sphere is that scarcity leads to weakness of voting position and respect. "Expand thy numbers," is the injunction each identity group responds to. Here I shall stay with the economist's case where per capita real incomes are hurt by abundance of numbers.

The simplest case is where the real national product,  $Q$ , is produced by the male and female adult labor forces,  $(M, F)$ , with the imputed real wage to each being determined by respective marginal products:

$$(8) \quad Q = Q[M, F] = Mq[F/M], \\ q' > 0 > q'';$$

$$W_m = \partial Q[M, F] / \partial M = q'[M/F];$$

$$W_f = \partial Q[M, F] / \partial F = \phi(W_m), \quad \phi' < 0.$$

*Warning:* a polar feminist model that supposed male and female inputs to be identical factors of production would make  $Q$  be  $c(M + F)$  and necessitate qualifying much of the following analysis.

If some people indulge a strong enough preference for sons, that will skew the adult sex ratio upward, tending to raise female wage rates and lower male rates. If women

are no less productive than men, in the sense that  $Q[M, F]$  is a symmetric function equalling  $Q[F, M]$ , then the female wage will exceed the male. Such an elevation in relative remuneration may serve *partially* to reverse the sex ratio imbalance.

One past reason for preferring sons may have been their superior earning power. Although daughters may perhaps be counted on to be more nurturing to you in your old age, sons may have the greater wherewithal to support you. Or the vanity of having the family name carried on may, under our patriarchal culture, be better served by sons. But now that the contrived scarcity of females raises their earnings, you have a new economic motive to indulge your preference for males less.

The new equilibrium, under the postulated symmetry of  $Q(M, F)$  and specified preference bias toward males, will be toward an excess of males but an excess limited by their induced impoverishment. The approach to the new equilibrium could involve successive over- and undershoots (as in the economist's "cobweb" model, where each person doesn't realize how many others are also doing the same thing). Or the approach may be gradual in the fashion of adaptive expectations. To keep up with the latest fad, we could even fabricate a rational expectations model: in it, people make a best guess about what the future development of the sex ratio will be; they take account of what is implied by this for relative earnings of their sons, daughters, grandsons and granddaughters, and in terms of this make their decisions; and, miracle of miracles, when all do this they together contrive what it is that they each expect.

Again we discover the possibility of a conflict between *group* (economic) selection and *individual* (economic) selection. If two unfriendly societies compete, the more prosperous one (Sparta) may eliminate the other. However, the  $M/(F + M)$  ratio that maximizes total  $Q$  (of Athens) may not be the sex ratio that best pleases its representative citizen. So each following self-interest may achieve the destruction of all.

*Observation:* genetic evolution can work to offset economic unbalancing of the sex ratio. To see this best, suppose  $p_f = p_m$  and

that only a subset of the population exercise gender control by abortion or otherwise. The groups who do and do not control gender need not be genetically defined. None the less, the dynamic Equation (7) could still apply, provided only we can suppose that when both parents belong to one group, so will the offspring. And that when the parents belong to different groups, half the offspring will be like one parent and half like the other.

What will follow? The Fisher logic will then apply to this non-particulate-inheritance setup. With what result? Our defined process of sociological selection will lead to a rebalancing of the sexes! Indeed, the process can lead to a serendipitous balancing if there are two groups, one favoring sons and one daughters: each can then get their heart's desire, with their numerical weightings evolving to keep the overall birth ratio in balance though no one is aware of the equilibrating process.

Mere deduction cannot prove anything definite about real world sex ratios. If we alter the razor's-edge symmetries of the Fisherine premises—admit nonrandom matings, etc., etc.—we of course can no longer deduce balanced sex ratios. Realistic economic models are not too likely to honor the precise “other things equal” premises implicit in the Fisherine deduction.

#### VIII. Finale

There is much territory between economics and biology that is still virgin ground. It will be tilled increasingly in the future. We should not be surprised if the first explorations are both crude and pretentious. Wisdom and maturity are the last settlers to arrive in pioneering communities.

#### REFERENCES

- Alchian, Armen A., “Uncertainty, Evolution and Economic Theory,” *Journal of Political Economy*, June 1950, 58, 211–21.
- Eshel, I., “Selection on Sex-Ratio and the Evolution of Sex-Determination,” *Heredity*, 1975, 34, 351–61.
- Fisher, R. A., *The Genetical Theory of Natural Selection*, Oxford: Clarendon Press, 1930.
- Friedman, Milton, *Essays in Positive Economics*, Chicago: Chicago University Press, 1953.
- Guttenberg, M. and Secord, P. F., *Too Many Women?—The Sex Ratio Question*, Beverly Hills: Sage Publications, 1983.
- Hamilton, W. D., “Extraordinary Sex Ratios,” *Science*, 1967, 156, 477–88.
- Henderson, L. J., *The Order of Nature: An Essay*, Cambridge: Harvard University Press, 1917.
- Karlin, S., “Mathematical Models, Problems, and Controversies of Evolutionary Theory,” *Bulletin (New Series) of The American Mathematical Society*, April 1984, 10, 221–73.
- Samuelson, Paul A., *Foundations of Economic Analysis*, Cambridge: Harvard University Press, 1947.
- Trivers, R. L. and Willard, D. E., “Natural Selection of Parental Ability to Vary the Sex Ratio of Offspring,” *Science*, 1973, 179, 90–92.
- Wilson, Edward, O., *Sociobiology*, Cambridge: Belknap Press, 1975.
- Winter, Sidney G. and Nelson, Richard R., *An Evolutionary Theory of Economic Change*, Cambridge: Belknap Press, 1982.
- Yellin, J. and Samuelson, P. A., “Genetic Fitness and the Sex Ratio: Natural Selection in Nonlinear Models of Population Growth,” unpublished NIH paper 1977.

# The New Economics of Labor Migration

By ODED STARK AND DAVID E. BLOOM\*

Research on the economics of labor migration has undergone an exciting and significant transformation during the past few years. At a theoretical level, migration research has expanded the domain of variables that seem to impinge upon and are affected by spatial labor supply decisions; it has highlighted the role of wider social entities and interactions within them in conditioning migration behavior; it has identified new linkages between migration as a distinct labor market phenomenon and other labor market and nonlabor market phenomena; and it has contributed to our understanding of the processes of economic betterment and development. At an empirical level, recent work on the economics of labor migration has confirmed the usefulness of old and well-established models of labor migration. It has also provided better estimates of key behavioral parameters, many of which are important ingredients in ongoing debates over public policies relating to migration. With such an impressive score, it is a wonder that more of the profession has not shifted into migration research. Perhaps this has to do with lack of information.

Our goal here is to summarize the actively evolving ideas, findings, and difficulties in the economics of labor migration. We do this mainly by illustrating selected theoretical and empirical developments which we believe to be on the frontier of research in this area. We also identify several new research topics

that comprise part of the next research frontier. Prior to proceeding with these tasks, we wish to point out that much of the more interesting recent research is associated with migration within and from developing economies. This situation might be partly explained by the fact that the impact of wage differentials on migration tends to be offset by unemployment compensation programs and other fiscal policies in the developed economies. The LDCs' scene thus constitutes a good migration research laboratory for studying migration in general.

## I. Theoretical Issues

Whereas owners of production inputs or commodities, such as bricks or bottles of wine, can ordinarily ship them away (so as to maximize profits or utility) while themselves staying put, owners of labor must usually move along with their labor. Furthermore, owners of labor have both feelings and independent wills. Indeed, most aspects of human behavior, including migratory behavior, are both a response to feelings and an exercise of independent wills. These simple observations divorce migration research from traditional trade theory as the former cannot be construed from the latter merely by effecting a change of labels.

People engage quite regularly in interpersonal income comparisons within their reference group. These comparisons generate psychic costs or benefits, feelings of relative deprivation or relative satisfaction. A person may migrate from one location to another to change his relative position in the same reference group, or to change his reference group. Membership in a low relative deprivation reference group may be well preferred to membership in a high relative deprivation reference group even if in the former a person's absolute income is lower. In general, a person who is more relatively deprived can be expected to have a stronger incentive to

\*Departments of Economics and Population Sciences, Harvard University, Cambridge, MA 02138 and Department of Economics, Bar-Ilan University; and Department of Economics, Harvard University, Cambridge, MA 02138, respectively. Comments by participants in the Harvard-MIT Research Seminar on Migration and Development are gratefully acknowledged. Bloom's research was supported by NIH Grant No. HD18844-02. A longer version of this paper, which includes references, is available from the authors upon request.

migrate than a person who is less relatively deprived. Moreover, a reference group characterized by more income inequality is likely to generate more relative deprivation and higher propensities to migrate. Note also that as particular individuals migrate, the relative deprivation perceived by nonmigrants may change, thereby creating second-round inducements to migrate. For example, if relative deprivation is gauged through a comparison with a reference group statistic such as average income, migration by low-income (i.e., relatively deprived) individuals will cause this statistic to increase and thereby induce migration by other individuals who become increasingly relatively deprived.

Not only can the migration behavior of individuals be expected to differ in accordance with their perceived relative deprivation, it can also be expected to differ according to their skill levels. This outcome results when the assumption of heterogeneous workers is paired with the assumption of imperfect skill information on the part of employers. To obtain some strong illustrative results, consider the following polar case. In a given profession, workers with skill  $S$  receive wages  $W_P(S)$  and  $W_R(S)$  from employers at  $P$  and  $R$ . Assume that skill follows a uniform distribution along a unit interval, that the functions  $W_P(S)$  and  $W_R(S)$  are nondecreasing and linear, and that  $S$  is known by  $P$  and  $R$  employers. Assume further that for low levels of  $S$ , say  $S < S^*$ ,  $W_P(S) > W_R(S)$ , whereas for  $S \geq S^*$  the reverse inequality holds. Clearly, the lowest-skilled workers will not wish to migrate. Assume now that  $R$  employers cannot observe the true skill level of individual  $P$  workers (i.e., that skill information is asymmetric), but that they know the distribution of  $S$  and will pay migrants from  $P$  a wage that is equal to the average productivity of the migrant group. The interior solution  $S^*$  now vanishes and is replaced by one of two corner solutions: there is either no migration at all, or there is migration by all. This result follows essentially because the highly skilled workers who migrate under perfect information may not do so if the pooled wage is too low. But if they do not, the pooled wage is lowered so that the next highly skilled group

also does not find it advantageous to migrate, and so on.

Just as it is clear that neither a brick nor a bottle of wine can *decide* to move between markets, so should it be equally clear that a migrant is not necessarily the decision-making entity accountable for his or her migration. Migration decisions are often made jointly by the migrant and by some group of nonmigrants. Costs and returns are shared, with the rule governing the distribution of both spelled out in an implicit contractual arrangement between the two parties. For example, one important component of the direct returns to the nonmigrating family from the migration of a family member are his or her remittances. Theory suggests the view, that empirical evidence seems to support, that patterns of remittances are better explained as an intertemporal contractual arrangement between the migrant and the family than as the result of purely altruistic considerations.

Theory also offers reasons for the migrant and the family to enter voluntarily into a mutually beneficial contractual arrangement with each other—rather than with a third party—and identifies conditions under which the contract is self-enforcing. Since the chosen contractual arrangement reflects the relative bargaining powers of the parties, this approach can also be used to generate empirically falsifiable predictions about remittance patterns, that is, that variables that enhance the bargaining power of the family and the importance of its support (such as a high-unemployment urban labor market) will *positively* influence the magnitude of migrant-to-family remittances. Note that this approach demonstrates the efficiency, flexibility, and what we might call the dynamic comparative advantage of the family. In other words, it does not view the family as an entity that is split apart as its independence-seeking younger members move away in an attempt to dissociate themselves from familial and traditional bondage, regardless of the negative externalities thereby imposed upon their families. Moreover, this approach shifts the focus of migration theory from individual independence (optimization against nature) to mutual interdependence (optimi-



zation against one another), that is, it views migration as a "calculated strategy" and not as an act of desperation or boundless optimism.

Risk handling provides another illuminating example in which a wider social entity is collectively responsible for individual migration. Clearly, the family is a very small group within which to pool risks. But the disadvantages of small scale may be made up by an ability to realize scale economies yet remain a cohesive group. Such scale economies are achieved by the migration of one or more family members into a sector where earnings are either negatively correlated, statistically independent, or not highly positively correlated with earnings in the origin sector. Again, as in the remittances example, the important point to note is that *both* parties are better off due to migration since, in this case, an exchange of commitments to share income provides coinsurance. Note, in addition, that just as it explains migration by part of the family, this example also accounts for nonmigration by the remainder.<sup>1</sup>

The nature of intragroup interaction could also help to explain features of the economic performance of migrants. To begin with, migrants often outperform the native born in the receiving economy. (We say more on this in Section II.) In addition, heavy reliance upon "network and kinship capital" is another prominent characteristic of migrant behavior patterns. The latter may explain the former quite readily in the context of an economy with a large number of agents whose transactions are governed by a prisoner's dilemma super game. Briefly, a migrant who offers to cooperate in his trade with *anyone* in the first game, whereas thereafter the choice in each game is that of the other agent in the previous game, will tend to be better off than a native who never behaves cooperatively, provided a sufficiently high propor-

tion of trades by migrants are conducted among migrants. This result provides an interesting explanation for the observation that new migrants are assisted by those who have migrated earlier; one good way of having a higher proportion of all trades conducted among migrants when there are few of them is to have additional migrants. The arrival of new migrants confers benefits upon the earlier migrants. It also suggests a resolution to the apparent inconsistency of altruistic behavior within a small group (say, a family) and selfish behavior within larger groups (say, a marketplace); the same strategy, viz, cooperate in the first game, thereafter reciprocate, is systematically applied throughout.

This appeal to strategic behavior may also be used to derive further migration-related insights. Consider, first, a not-atypical village economy in an *LDC* where farming landlords are in an oligopsonistic position with respect to the determination of wages and employment. Through collusion, the farmers can increase their profits. However, labor migration can constitute a credible counter-strategy to this possibility, provided that, from time to time, some undertake it. Note that once again, migration confers benefits upon those who stay behind, in addition to those associated with a leftward shift in the supply curve of labor. Second, consider the case of employers who, in static and dynamic contexts alike, are better off with a larger labor pool than with a smaller labor pool. Since a large labor pool can be developed by cultivating an image of worker success, it might be worthwhile for employers to create high-paying jobs in order to attract more migrants. As long as a large number of workers have the belief that high-paying employment can be obtained, or that it is worth waiting for, a migratory response will be produced. High "institutionally determined" wages in urban labor markets in *LDCs* are thus not necessarily externally imposed upon reluctant employers by government legislation and trade unions. Instead, they may result from endogenously determined strategies designed to maximize profits in dynamic settings. Also, generating few very high-paying jobs and heavily advertising, so to speak, the rewards associated with them may help

<sup>1</sup>The insurance attribute of migration applies to the individualistic case too. For example, just as general human capital provides self-insurance, so does migration in conjunction with specific human capital. Thus, in easing risk bearing associated with investment in specific human capital, migration facilitates such investment thereby conferring efficiency gains.

to maintain a large labor pool in the presence of high levels of unemployment. This strategy will tend to confuse migrant calculations, which may suggest that expected urban income is less than rural income. Thus, high-paying jobs might also be created *in response* to high levels of unemployment rather than preceding them and bringing them about.

Since the endowments and preferences of economic agents are always heterogeneous in practice, selectivity, as such, in response to a given set of prices and opportunities and changes in it, by way of migration or otherwise, is quite obvious. In many cases, whether migration selectivity prevails is not as interesting as the extent to which the migration response diffuses. Indeed, migration can be looked upon as a process of innovation adoption and diffusion. As time goes by, what proportion of a given group of *potential* migrants have migrated? To illustrate, assume there are a number of migration destinations and that there is some prior belief that one particular destination is better than the others. In this setting, the experience of actual migrants provides valuable information that presumably reduces future uncertainty of the remaining pool of potential migrants. Under these circumstances, the most interesting research issues relate to the determination of the speed of adoption of migration as an innovation and the characteristics associated with the delay in the adoption of the innovation (rather than whether it takes place), that is, why are some individuals quicker to migrate than others? For the case of rural-to-urban migration in *LDCs* where, if history were to repeat itself, most rural people will end up as migrants, such an approach seems particularly appropriate. Note that as with a demonstration effect in the case of innovation adoption, a stock of past migrants at a given destination (particularly a large stock) represents evidence that might lead to an upward revision of beliefs that migration is a worthy investment. Moreover, the impact of migration upon the society from which it takes place is now stage-specific. Thus, the divergence of views concerning the consequences of migration (for example, its impact upon the distri-

bution of income by size) can partly be attributed to the simple fact that the underlying observations are made at distinct stages of the diffusion process.

## II. Empirical Considerations

Recent empirical research on the economics of labor migration has benefited a great deal more from the development of new econometric techniques than from new theoretical ideas. The techniques that have substantially improved our ability to use micro data sets in the estimation of relatively standard models of labor migration include techniques for the analysis of qualitative dependent variables, techniques that correct for sample selection bias, and techniques for the analysis of longitudinal and pseudo-longitudinal data. At the micro level, most empirical studies have attempted to test simple microeconomic models of migration according to which individuals (or families) make locational decisions primarily by comparing their income opportunities at alternative locations. The key feature of recent studies of this type is their focus on the estimation of structural, as opposed to reduced-form, models of the migration decision. In the past, a major problem that made the estimation of such models difficult was the absence of data on the wages that particular individuals would receive at two or more locations at the same point(s) in time. In other words, survey data sets typically provide researchers with information on the wages received by individuals at their residential location at the time of the survey, their migrant or nonmigrant status at that location, and selected individual characteristics (for example, age, education, and marital status). To the extent that particular *unobserved* characteristics of individuals are rewarded differently at different locations, the average wage of individuals (conditional on their observed characteristics) at location *A*, who migrated there from location *B*, will provide a biased estimate of the wage that individuals who remained at location *B* would receive if they moved to location *A*.

Largely as a result of advances in the statistical analysis of selected samples, however, we now have fairly simple methods that

we can use to test and correct for the bias associated with this unobserved wage problem. To date, estimates of these structural models of labor migration uniformly support the hypothesis that individuals respond to income incentives in making decisions to migrate. However, further application of these models is desirable, using different data sets and more carefully formulated and tested empirical specifications. It would be interesting to examine whether the strength of the migration response to wage differentials decreases over time, while the response to variables such as relative deprivation increases. We would also like to point out that longitudinal data may prove particularly useful in analyzing the determinants of migration, insofar as they permit a distinctly different approach to the problem of sample selection (i.e., longitudinal data permit researchers to control more directly for unobserved variables that affect wages and that are correlated with the migration decision).

Furthermore, much empirical research has been conducted on the labor market progress of migrants, with special attention paid to the behavior of international migrants. To date, most studies of this topic have involved the estimation of cross-sectional wage equations in which "years since migration" is entered as an independent variable and its coefficient is interpreted as a measure of migrant progress. Typically, these studies find that migrant workers earn less than native-born workers with similar characteristics during the first few years after migration but more thereafter. It has been suggested, however, that this longitudinal conclusion, based on analyses of cross-sectional data, may be an artifact of either the declining quality of migrant labor over time (i.e., a vintage effect) or the outmigration of the least successful migrants. In view of the contradictory nature of extant empirical conclusions, and given the academic and policy importance of this issue, additional research on the pace of migrants' labor market progress is clearly needed. Further analysis of longitudinal data on migrant earnings would also be helpful.

In addition to the two focal points for empirical work discussed above, there are four other areas that empirical economists

have touched upon and which we think should receive further attention. The first of these areas involves estimation of the macroeconomic effects of migration. There is a surprising lack of empirical work on the effects of labor migration on wages and employment in net-sending and net-receiving locations, especially for different types of labor (for example, skilled and unskilled labor). Further work on this topic would be of interest, perhaps involving estimation of the wage and employment effects of migration in the context of well-defined structural models of equilibrium and disequilibrium labor markets. Analysis of the distributional impacts of migration and the degree of substitutability between international and internal migration in the process of labor market adjustment would also be helpful.

Second, the microeconomic and macroeconomic relationships between aging and labor migration are topics which have received only scant and indirect empirical attention (for example, age is usually a right-hand side variable in microeconomic studies of migration decision making). Indeed, empirical evidence strongly suggests that older workers are less mobile than younger workers. This finding is quite plausible for a variety of reasons relating to the differential preferences and opportunities of older and younger workers. It therefore seems likely that workforces in many low-fertility countries will show a declining propensity to respond to exogenous economic change by migration as they age over the next two decades. Thus, to the extent that mobility is one of the key requirements for economic efficiency, it would be useful to know more about the extent to which the aggregate migration behavior of a population is influenced by its age distribution and the underlying bases for this relationship. Such information could be very helpful in debates over public policies that provide incentives to migrate.

The third topic that deserves further empirical attention is the migration behavior of dual-earner families. In its most general form, this issue relates to the broader one of the appropriate unit of analysis for studying migration behavior to which we alluded in Section I, that is, the individual or the family.

# New Evidence on the Timing and Spacing of Births

By JAMES J. HECKMAN, V. JOSEPH HOTZ, AND JAMES R. WALKER\*

This paper is a first progress report of an ongoing empirical study of the determinants of life cycle fertility (see our 1985 paper for a more complete report). At this stage, our analysis is decidedly empirical. Unlike many other areas of knowledge in economics and social science, there are few widely accepted or carefully confirmed "stylized facts" in fertility dynamics to guide economic model builders. This vacuum has inhibited the successful development of economic models of fertility dynamics. The main objective of the early stage of our work is to codify the "facts" in a coherent statistical framework that provides the duration data analogue of the conventional simultaneous equations model.

The starting point for our analysis is the demographic literature. This literature finds that the age at marriage, the occurrence of births inside or outside of marriage, the age of the first birth and/or the durations of previous birth intervals significantly affect the timing of subsequent births over the life cycle. Studies of *marital fertility* by L. Coombs and R. Freedman (1970), L. Bumpass, R. Rindfuss, and R. Janosik (1978), J. Trussell and J. Menken (1978), F. Finnas and J. Hoem (1980), and Hoem and R. Selmer (1984) find that the younger the woman's age at marriage, the more rapid the pacing of subsequent fertility, and that those who begin childbearing early in their reproductive careers subsequently have children more rapidly.

Recent cross-country comparisons of data from the World Fertility Survey have found that the timing of marriage and the lengths

of prior birth intervals directly affect the spacing of subsequent life cycle fertility. Such results have lead G. Rodriguez et al. to conclude that "birth interval lengths depend little upon birth order, but far more upon the length of the previous interval" and that the human "reproductive process can be encapsulated as an engine with its own inbuilt momentum whereby early behavior and socio-economic differences fundamentally determine (along with ageing and secular variation) the remainder of the childbearing experience" (1984, p. 5). Such a view suggests: 1) that the age of entry into marriage (or marriage-like unions, such as cohabitation) and/or the entry into parenthood are the crucial determinants of life cycle fertility and that the variation in completed fertility across the population comes primarily in these initial decisions; and 2) that subsequent childbearing is largely determined by initial events and by the lengths of preceding birth intervals. Another widely held view is that after the first birth interval, subsequent birth intervals are biologically determined until the birth process terminates.

One objective of our study is to investigate the robustness of these findings in other data sets. For a sample of Swedish women described below, we find that the conventional demographic view of the fertility process is largely confirmed on fresh data. The main objective of our study, however, is to probe the empirical findings to discriminate between "structural" and "spurious" explanations for the observed empirical regularities. As is the case in any empirical study, only a subset of plausible determinants of life cycle fertility is observed. Evidence that lagged dependent variables affect current values of those variables or evidence that initial outcomes of a life cycle process affect later outcomes may be due to serial correlation in unobservables.

In the context of fertility, one might expect a woman who desires children or who

\*The University of Chicago and The Economics Research Center/NORC, Chicago, IL 60637. We thank George Yates for his help. This research was supported by Grant No. HD 19226 from the National Institute of Child Health and Human Development.

has a comparative advantage at the activity of child rearing might marry earlier, have children at an earlier age, and invest in household capital relative to a woman with a stronger motivation for market activity. Failure to control for unobserved propensities and skills may explain the regularities noted above. On the other hand, the observed empirical regularities may be the outcome of genuine behavioral influences of past events on current constraints or preferences. For example, early childbearing may cause a woman to invest less in market skills and more in nonmarket skills, leading to higher fertility rates for such women. Our study attempts to shed light on the issue of the causal significance of these demographic empirical regularities. A main conclusion is that controlling for unobservables in a robust nonparametric fashion vitally affects the sign and statistical significance of the estimated effect of early life cycle events on subsequent fertility outcomes. Some of the stylized facts of the demographic literature are not robust to controls for unobservables but others remain.

We find that for a variety of empirical specifications of the life cycle birth process and for a variety of samples that do and do not condition on marital status, in models that do not control for unobservables, the longer a preceding birth interval the longer the subsequent one. Thus, our data exhibit the "engine of fertility" phenomenon noted by demographers. Controlling for unobservables, the "well-noted empirical regularity" either vanishes or reverses in sign. For married women, it vanishes entirely. For a sample of all women in a model that controls for marital status as a covariate, controlling for unobserved heterogeneity produces a "reverse engine of fertility" phenomenon: the longer the preceding birth interval the *shorter* the subsequent one. For a sample of married women, the importance of the age of marriage on the spacing of birth intervals is considerably reduced in size and statistical significance. For a sample of all women, controlling for unobserved variables eliminates the impact of age at marriage on all but the final birth transition.

## I. The Data

The data used in this analysis are from a survey of retrospective interviews conducted by the Swedish National Central Bureau of Statistics in 1981. The survey asked for complete cohabitational, marital, childbearing, and work histories as well as background information of almost 5,000 Swedish women randomly selected from five recent cohorts of women of all marital statuses. (For further description of the data set, see Hoem and B. Rennermalm, 1982.) We have restricted our analysis to a sample of women who were born between 1941–45. For this cohort, we had useable data for 990 women. In order to analyze the effects of sample conditioning in previous studies of marital fertility, we also estimated some models described below on a subset of women from this cohort who: (a) were married at least once prior to 1981, (b) experienced their first "union" (i.e., marriage or consensual union) before age 25, and (c) did not have a birth prior to their first marriage. There are 570 such women.

## II. The Model

For the sake of brevity, we only sketch the statistical model used to perform the empirical work reported below. A full description of the model is available in our companion paper. The main building block is the multistate hazard rate:

$$(1) \quad h_{ij}(t_{ij} | \mathbf{x}'\beta_{ij}, c_{ij}\theta),$$

where  $i$  denotes the origin state and  $j$  denotes the destination state,  $t_{ij}$  is the duration in state  $i$  which exits into state  $j$ ,  $\mathbf{x}$  is a vector of (possibly time varying) observed variables that may include lagged durations or occurrence times of previous events,  $\beta_{ij}$  is a vector of associated coefficients,  $\theta$  is a scalar unobservable, and  $c_{ij}$  is a transition-specific factor loading. Specifying the functional form for the hazard and using the nonparametric maximum likelihood estimation procedure of Heckman and B. Singer (1984b), it is possible to consistently estimate the parameters of the hazard (including the

$c_{ij}$ s) and the population distribution of the unobservables. Allowing the  $h_{ij}(t_{ij})$  to be a nontrivial function of duration  $t_{ij}$  allows for spell-specific duration dependence. For a further discussion of multistate models, see Heckman and Singer (1984a).

A birth process specializes the model to transitions among successive birth states,  $i = 0, \dots, I$ , where  $I$  is the maximum number of births. In this specification,  $j = i + 1$  for each  $i$ . In the restricted birth process model, transitions among marital states are treated as changes in exogenous variables (changes in elements of  $\mathbf{x}$ ). Note that each duration is permitted to be governed by different parameters to explicitly allow for different birth order duration dependence and effects of the  $\mathbf{x}$ 's. Below, we test whether or not the set of parameters are significantly different for the intervals between the second and third births to examine the finding in Rodriguez et al. of no-birth-order effects on higher-order birth intervals. Finally, we note that because women are followed from the age of menarche, there is no problem of initial conditions. (See Heckman and Singer, 1984a, for a discussion of this problem.) In a more general multistate process, it is possible to introduce transitions among birth and nonbirth states. Specifically, in our companion paper, we permit transitions from the single or cohabiting state to the states of marriage and cohabitation, transitions among all three of these states and transitions to births from all possible states.

The empirical results reported below are based on a flexible model with linear and quadratic duration terms:

$$(2) \quad h_{ij}(t_{ij}|\mathbf{x}'\beta_{ij}, c_{ij}\theta) = \exp(\gamma_{0ij} + \gamma_{1ij}t_{ij} + \frac{1}{2}\gamma_{2ij}t_{ij}^2 + \mathbf{x}'\beta_{ij} + c_{ij}\theta).$$

### III. Empirical Results

The variables utilized in our analysis are defined as follows: *DURATION* = number of months/100 spent in current spell; *DURATION*<sup>2</sup> = the square of the number of months/100 spent in current spell; *AGE-*

*UNION* = the woman's age (in months/100 since her thirteenth birthday) at which she first entered either marriage or a consensual union; *AGEMAR* = the woman's age (in months/100 since her thirteenth birthday) at which she was married, if she married, and 0 otherwise; *EVERMAR* = a dummy variable = 1 at the age she gets married and = 0 prior to that age; *AGECOH* = the woman's age (in months/100 since her thirteenth birthday) at which she first cohabitated, if she cohabitated, and 0 otherwise; *EVERCOH* = a dummy variable = 1 at the age she begins cohabiting and = 0 prior to that age; *BIRTHDUR1* = the duration in months/100 of the first birth (for the ever married subsample it equals age at first birth—age at first union and for the full sample it equals number of months/100 since her thirteenth birthday); *AGE1STBIR* = age at first birth in months/100 since her thirteenth birthday; *BIRTHDUR2* = the duration in months/100 from first birth until the second birth; *LFP* = a time-varying variable = 1 if the woman is working at the current duration and 0 otherwise; *EDUC* = a time-varying = 1 if the woman is in school at the current duration and 0 otherwise; *URBAN* = a dummy variable = 1 if the woman grew up in an urban area of Sweden and 0 otherwise; *WHITECOL* = a dummy variable = 1 if the woman's father had a white-collar occupation when she was growing up and 0 otherwise; *UNIV* = a dummy variable = 1 if the woman attended a university by the time of the interview and 0 otherwise.

Table 1 reports empirical estimates of the parameters of parity specific hazard functions for *ever married* women. Panel A reports estimates for a statistical model that does not control for unobservables; panel B reports the estimates controlling for unobservables.

Panel A is largely consistent with the engine of fertility story. The age of first union (marriage or cohabitation) is positively associated with the age of first birth and negatively associated with the durations of subsequent births. The length of the previous birth interval is positively associated with the length of the subsequent interval. (This is so

TABLE 1—ESTIMATES FOR BIRTH  
PROCESS HAZARD RATES

Variable	AGEUNION to First Birth	First Birth to Second Birth	Second Birth to Third Birth
A: Not Controlling for Heterogeneity			
CONSTANT	1.77 (0.154)	1.05 (0.245)	2.01 (0.369)
DURATION	2.15 (0.411)	12.00 (0.650)	4.84 (0.916)
DURATION <sup>2</sup>	-4.91 (0.743)	-19.8 (1.030)	-5.92 (1.240)
AGEUNION	0.406 (0.118)	-0.349 (0.160)	-0.613 (0.375)
BIRTHDURI		-0.663 (0.166)	
BIRTHDUR2			-3.64 (0.567)
LFP	-3.82 (0.142)	-2.62 (0.180)	-3.10 (0.368)
EDUC	-3.14 (0.264)	-2.50 (0.574)	-1.94 (0.760)
URBAN	0.476 (0.066)	0.109 (0.074)	-0.244 (0.148)
WHITECOL	-0.055 (0.066)	-0.093 (0.080)	0.209 (0.170)
UNIV	0.359 (0.122)	1.14 (0.129)	1.250 (0.252)
Log Likelihood		749.8	
B: Controlling for Heterogeneity			
CONSTANT	1.04 (0.205)	0.350 (0.383)	-5.18 (1.120)
DURATION	2.70 (0.452)	17.33 (0.945)	7.05 (0.976)
DURATION <sup>2</sup>	-4.98 (0.793)	-22.20 (1.266)	-6.94 (1.282)
AGEUNION	0.697 (0.160)	-0.250 (0.284)	-1.07 (0.455)
BIRTHDURI		0.234 (0.315)	
BIRTHDUR2			1.23 (1.003)
LFP	-4.00 (0.147)	-4.21 (0.205)	-4.44 (0.370)
EDUC	-3.34 (0.276)	-3.09 (0.708)	-2.56 (0.797)
URBAN	0.399 (0.079)	0.123 (0.131)	-0.296 (0.212)
WHITECOL	-0.031 (0.081)	-0.150 (0.141)	0.125 (0.242)
UNIV	0.217 (0.150)	1.37 (0.230)	2.15 (0.377)
Factor	1.50	4.58	8.16
Loading	(0.145)	(0.330)	(1.042)
Log-Likelihood		942.6	

Note: Sample: Continuously married Swedish women, birth cohort 1941-45 ( $N=570$ ); asymptotic standard errors are shown in parentheses.

because a negative effect of preceding duration on a current hazard means the transition rate to the next child is smaller for the next duration and hence the interbirth interval tends to be larger.) Background variables such as father's occupational status play a minor role.

Controlling for unobserved heterogeneity using a nonparametric maximum-likelihood procedure causes the effect of length of the preceding birth interval on current transition rates to vanish. In fact, for the latter two transitions, the estimated effect of lagged births switches sign. The effect of age at marriage on the estimated transition rates becomes *stronger* for the transition to the first birth and for the transition from the second to the third birth. The fact that unobservables are empirically important is reflected in the great improvement in the fit of the model to the data as measured by the log likelihood.

Table 2 reports empirical estimates for a sample of all women irrespective of their marital history. The new explanatory variables reported in that table are introduced to control for marital history. Such "controls," while traditional, are *ad hoc*. In our companion paper, we estimate a multistate model that breaks out transitions from each marital state to fertility and allows for endogenous transitions among marital and fertility states.

Controlling for heterogeneity in a model estimated on this sample of women causes the effect of age at marriage to vanish from the model (except for the last transition) and causes the effect of the length of the preceding interval on the transition rate to the next birth to reverse sign. *For the full sample of women, the engine of fertility works in reverse.* Controlling for unobservables, a long first birth interval leads to a short second birth interval and a long second birth interval leads to a short third birth interval. These results are consistent with a fixed target model of fertility in which a delay in the arrival of one child is compensated for by an acceleration of the rate of arrival of the next child.

The hazard functions reported for both samples control for the woman's current labor force participation and school atten-

TABLE 2—ESTIMATES FOR BIRTH PROCESS HAZARD RATES

Variable	Age 13 to First Birth	First Birth to Second Birth	Second Birth to Third Birth
A: No: Controlling for Heterogeneity			
CONSTANT	1.43 (0.057)	1.49 (0.226)	0.062 (0.418)
DURATION	2.67 (0.258)	9.34 (0.470)	3.73 (0.703)
DURATION <sup>2</sup>	-2.01 (0.163)	-14.22 (0.686)	-4.67 (0.929)
AGEMAR	0.316 (0.111)	0.368 (0.156)	-0.814 (0.259)
EVERMAR	1.08 (0.136)	0.128 (0.223)	1.05 (0.472)
AGECOH	0.121 (0.117)	0.004 (0.131)	-0.232 (0.323)
EVERCOH	0.306 (0.127)	-0.073 (0.139)	0.247 (0.311)
AGE1STBIR		-0.646 (0.134)	
BIRTHDUR2			-2.06 (0.349)
LFP	-4.14 (0.106)	-2.88 (0.131)	-2.70 (0.234)
EDUC	-4.28 (0.218)	-2.64 (0.346)	-2.02 (0.619)
URBAN	0.225 (0.047)	0.040 (0.057)	-0.036 (0.118)
WHITECOL	-0.183 (0.048)	-0.118 (0.061)	-0.136 (0.139)
UNIV	0.388 (0.083)	1.23 (0.099)	1.36 (0.190)
Log Likelihood		748.0	
B: Controlling for Heterogeneity			
CONSTANT	2.43 (0.096)	3.61 (0.322)	1.32 (0.522)
DURATION	3.46 (0.328)	14.00 (0.592)	6.85 (0.783)
DURATION <sup>2</sup>	-2.13 (0.203)	-16.9 (0.790)	-7.10 (1.010)
AGEMAR	-0.0005 (0.151)	-0.087 (0.237)	-0.649 (0.371)
EVERMAR	1.25 (0.171)	1.03 (0.310)	1.72 (0.635)
AGECOH	0.232 (0.149)	-0.223 (0.233)	-0.119 (0.437)
EVERCOH	0.272 (0.157)	0.429 (0.243)	0.426 (0.413)
AGE1STBIR		0.360 (0.188)	
BIRTHDUR2			0.771 (0.385)
LFP	-4.54 (0.113)	-4.44 (0.140)	-3.89 (0.241)
EDUC	-4.47 (0.226)	-3.47 (0.391)	-2.47 (0.676)
URBAN	0.182 (0.068)	-0.005 (0.099)	0.126 (0.167)
WHITECOL	-0.144 (0.074)	-0.199 (0.109)	-0.505 (0.201)
UNIV	0.110 (0.127)	1.08 (0.159)	1.88 (0.265)
Factor Loading	-2.16 (0.120)	-4.58 (0.229)	-8.96 (0.638)
Log-Likelihood		1117.9	

Note: Sample: All Swedish women, birth cohort 1941-45 ( $N = 990$ ); asymptotic standard errors are shown in parentheses.

dance statuses as covariates. While some of the studies cited above include these variables, it may be argued that they are endogenous variables. In results not reported here, we reestimate the model excluding these two variables. Our conclusions are not changed when they are omitted.

The main point to extract from our analysis is the fragility of the "empirical regularities" to the introduction of unobservables. The "stylized facts" currently offered by demographers do not recognize the importance of unobservables. It would be unfortunate if economists were to take as their mission the formulation of economic models to explain phenomena generated by the simple fact that people are different.

An encouraging feature of these empirical results is that, for the Swedish data, we decisively reject the "biological determination" view that higher parity hazards are identical. Controlling for heterogeneity that could lead to spurious differences in parity specific hazards even if the conditional hazards (given the unobservable) were equal across parities, we reject the hypothesis that the transition rate from the first to the second birth is identical to the transition rate from the second birth to the third birth. The  $\chi^2$  statistics for the hypothesis that the corresponding coefficients of these two transitions are the same are: 581.9 with 11 degrees of freedom for the ever married sample of women and 351.0 with 14 degrees of freedom for the full sample of women.

#### IV. Conclusions

Our results demonstrate the importance of controlling for unobservables in the analysis of life cycle birth processes. We conclude this paper by noting an additional issue that is also of potential importance in the analysis of fertility data and which we investigate in our companion paper.

Variations in the sample selection rules used can affect the inference from estimates, such as those presented above, in a nontrivial way. All of the studies cited above are for marital fertility. None correct for the sample selection bias that arises from using such behaviorally conditioned samples. Correc-



tions for selection have been shown to be empirically important in many studies (see Mark Killingsworth, 1983; James Smith, 1980). The appropriate generalization of methods for dealing with selection bias in a dynamic setting is a multistate duration model. In our companion paper, we introduce transitions among birth and nonbirth states using a general multistate process. Specifically, we permit transitions from the single or cohabiting state to the states of marriage and cohabitation, transitions among all three of these states and transitions to births from all possible marital states. We examine the impact of correct conditioning on estimated marital fertility hazard rates, utilizing a multistate specification. Explicitly accounting for the transitions among these life cycle states in estimation of the parameters of the fertility processes corrects for the selection on marital status.

#### REFERENCES

- Bumpass, L., Rindfuss, R. and Janosik, R., "Age and Marital Status at First Birth and the Pace of Subsequent Fertility," *Demography*, February 1978, 12, 75-86.
- Coombs, L. and Freedman, R., "Premarital Pregnancy, Childspacing and Later Economic Achievement," *Population Studies*, November 1970, 24, 389-412.
- Finns, F. and Hoem, J., "Starting Age and Subsequent Birth Intervals in Cohabitation Unions in Current Danish Cohorts, 1975," *Demography*, August 1980, 17, 275-95.
- Heckman, J. and Singer, B., (1984a) "Econometric Duration Analysis," *Journal of Econometrics*, January 1984, 24, 63-132.
- \_\_\_\_\_, and \_\_\_\_\_, (1984b) "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, March 1984, 52, 271-320.
- \_\_\_\_\_, Hotz, V. J. and Walker, J., "The Determinants of Birth Intervals: A Multistate Approach," unpublished manuscript, University of Chicago, January 1985.
- Hoem, J. and Rennermalm, B., "Cohabitation, Marriage, and First Birth among Never-Married Swedish Women in Cohorts Born 1936-1690," unpublished manuscript, University of Stockholm, December 1982.
- \_\_\_\_\_, and Selmer, R., "The Negligible Influence of Premarital Cohabitation on Marital Fertility in Current Danish Cohorts, 1975," *Demography*, May 1984, 21, 193-206.
- Killingsworth, M., *Labor Supply*, Cambridge, England: Cambridge University Press, 1983.
- Marini, M. and Hodsdon, P., "Effects of the Timing of Marriage and First Births on the Spacing of Subsequent Births," *Demography*, November 1981, 18, 529-48.
- Rodriguez et al., G., "A Comparative Analysis of the Determinants of Birth Intervals," *Comparative Studies*, -No. 30, London: World Fertility Survey, April 1984.
- Smith, J., "Introduction," in his *Female Labor Supply: Theory and Evidence*, Princeton: Princeton University Press, 1980.
- Trussell, J. and Menken, J., "Early Childbearing and Subsequent Fertility," *Family Planning Perspectives*, July/August 1978, 10, 209-18.

## PERSPECTIVE ON THE EXTERNAL DEBT SITUATION<sup>†</sup>

### International Debt: From Crisis to Recovery?

By WILLIAM R. CLINE\*

Since Mexico temporarily suspended payment in August 1982, the debt crisis has threatened the international financial system and crippled growth in developing countries. Emerging evidence and projection analysis suggest that the problem can be managed, however, if satisfactory growth in the industrial countries can be maintained.

#### I. Origins, Systemic Risk, and Emergency Response

From 1973 to 1982, external debt of non-oil developing countries rose by \$500 billion. Of this amount approximately \$260 billion may be attributed to the exceptional rise in oil prices. Global recession in 1981–82 added another \$100 billion through declines in the terms of trade and reduced export volume; and the excess of real interest rates in this period over historic averages cost them another \$40 billion. External shocks thus accounted for a major portion of the debt crisis. Domestic factors also contributed, especially overvalued exchange rates and inadequate domestic interest rates that caused capital flight (Mexico, Venezuela, and Argentina).

Because debt tends to grow at the interest rate (by "inheritance" from past debt, unless the country runs a trade surplus to pay interest), exports need to grow at least this fast or else the burden of debt relative to exports increases. From 1973 to 1980, international interest rates (including a typical spread above LIBOR) averaged 10 percent, while

nominal export growth from non-oil LDCs averaged 21 percent. But in 1981–82, the interest rate rose to 16 percent while export growth averaged 1 percent. The debt problem may usefully be viewed as the consequence of the shift from low or negative real interest rates in the inflationary 1970's to the high real interest rates of the early 1980's, aggravated by declining real exports during the global recession.

In mid-1982, the nine largest U.S. banks had loans outstanding to developing countries and Eastern Europe amounting to 280 percent of their capital, and most had over 100 percent of capital in loans to just Brazil and Mexico. Large losses on LDC debt could cause technical insolvency in these banks. Attempts to bail out the banks could prove inflationary, while sharp cutbacks in bank capital would mean contractionary pressure as they cut back lending to maintain capital-asset ratios. Although some monetarists have argued that even the large banks could fail without adversely affecting the system if the Fed would maintain steady growth in the money supply, this approach would amount to a global roll of the dice.

Not being gamblers, policymakers responded to the debt crisis with emergency financial packages composed of: 1) a country adjustment program under IMF auspices; 2) continued new lending by banks; and 3) financial support from the IMF, central banks, and multilateral lending institutions.

#### II. Modeling Debt Viability

The premise underlying this policy response was that the debt crisis was one of illiquidity, requiring temporary financing, rather than one of insolvency, involving outright loss of a considerable portion of principal. To examine this question, in the spring of 1983, I developed a projection model for

<sup>†</sup>*Discussants:* Jeffrey Frankel, University of California-Berkeley; Basil G. Kavalsky, World Bank; Henry Wallich, Board of Governors of the Federal Reserve System.

\*Institute for International Economics, 11 Dupont Circle, N.W., Washington, D.C. 20036. References not otherwise specified may be found in my 1984 book.

balance of payments and debt of the 19 largest debtor countries (see my 1983 study).

The model relates exports of the debtor country to OECD growth and the country's real exchange rate. On the basis of a statistical relationship between OECD growth and aggregate non-oil imports, export volume from developing countries is estimated to decline by 3 percent when OECD growth is zero and to increase above this rate by 3 percentage points for each percentage point of OECD growth. The marginal elasticity of exports is 3, and the average elasticity at a typical OECD growth rate of 3 percent is 2. Export prices rise beyond general inflation when OECD growth increases (terms of trade are procyclical), based on relationships estimated for individual countries with data for the past two decades. On average, a 1 percent rise in OECD growth causes approximately 1.5 percent rise in real export price in the current year and a similar rise in the following year. An elasticity of 0.5 is used for the response of exports to the country's real exchange rate.

The model applies a unitary elasticity of real imports with respect to the country's domestic growth, plus an elasticity of three for changes in that growth rate (cyclical elasticity). The elasticity of imports with respect to real exchange rate is  $-0.6$ . Interest payments depend on the previous year's debt as applied to the current year's international interest rate (LIBOR) and a spread. Oil trade is fixed in volume, with values depending on assumed oil prices.

All traded goods (except oil) rise in price, additionally, at the international inflation rate. Also, if the dollar depreciates internationally, dollar prices of traded goods rise (otherwise these goods would command fewer resources internationally because of a change in the dollar-yen or dollar-mark rate) by 80 percent of the depreciation on the basis of trade prices in recent years. Dollar depreciation boosts the dollar value of non-oil exports relative to debt, which is mainly denominated in dollars.<sup>1</sup>

Using this model, my estimates in mid-1983 indicated that the debt problem was indeed one of illiquidity rather than insolvency. Under central expectations for international economic variables (and politically acceptable growth rates in debtor countries), the projections showed that most major countries would show substantial improvement in balance of payments and relative debt burden, and that by the late 1980's, debt-export ratios would be back to levels previously associated with creditworthiness. A return to higher OECD growth would increase export volume and prices, an eventual easing of interest rates would moderate interest payments, and a decline in the dollar from its seriously overvalued level would raise the dollar value of the export base. However, the analysis also indicated that a critical threshold of  $2\frac{1}{2}$  to 3 percent was required for OECD growth to avoid stagnation or severe deterioration in external deficits and debt-export ratios for debtor countries.

Even with gradual recovery from the debt crisis, however, there remains a difficult interim period during which major debtors will not yet have restored creditworthiness to levels necessary for normal capital market access. Where will financial flows come from in this period? Here I have suggested a model of "involuntary lending." In this model, although a country's creditworthiness is too weak for new creditors to risk lending, existing lenders with current exposure will be prepared to make modest new loans to shore up the quality of existing loans. Existing lenders will extend new credit as long as the expected benefit, the reduction in probability of default multiplied by the amount of outstanding exposure, exceeds the expected cost of new lending, the terminal default probability multiplied by the amount of the new loan.

---

would worsen because of the decline in U.S. imports from debtor countries. However, if the country pegs its exchange rate with weights proportionate to partner trade shares, and if trade elasticities are symmetrical, a decline in the dollar will cause increased exports to Europe and Japan that offset any decrease in exports to the United States. Only oil exporters, with oil prices insensitive to the dollar, are likely to be adversely affected by dollar depreciation.

<sup>1</sup>Some appear to believe that if the dollar were to depreciate from its current high level, the debt problem

This model predicts a relatively robust process of continued lending at modest rates. However, it is subject to breakdown because of the free-rider problem. Smaller banks are likely to judge their own actions to have little effect on default probability—even though in the aggregate their actions will affect the outcome.

In a historic departure, the IMF set as a precondition for its own lending that the banks as a group provide significant new lending. Central banks also apparently twisted some arms to ensure participation by smaller banks. These actions helped ensure burden sharing of the public good of increased lending. However, considering that up to two-thirds of bank exposure is held by perhaps 50 to 100 banks internationally, it is likely that a critical mass of banks would have found it in their own interest to provide additional lending even with less official intervention.

So far the mechanism of involuntary lending has held up relatively well. From June 1982 to June 1984, the exposure of U.S. banks in 6 major Latin American debtor countries rose by 9 percent, despite significant reductions in exposure in Argentina and Venezuela by banks other than the largest 24 (Federal Financial Institutions Examination Council, 1984). This rate is approximately equal to the 5 percent annual rate frequently cited as necessary to bridge the period of involuntary lending. Globally, 5 percent growth in exposure corresponds to approximately \$20 billion annually in new bank lending (down from \$50 billion in 1981).

For their part, debtor countries have cooperated and carefully avoided an aggressive debtors' cartel. Even though their interest payments exceed new borrowing, they have little incentive to default, because of immediate adverse consequences (drying up of trade credit, possible seizure of export shipments) and long-term damage to their credit reputation.

### III. Recent Evidence

The emerging evidence in 1983-84 has tended to confirm the analysis that the debt problem is one of illiquidity and subject to

improvement as international recovery takes place. In 1983, the 19 largest debtor countries (accounting for three-fourths of bank debt) cut their external current account deficits from \$56 billion to \$23 billion (an even sharper improvement than in my original projections). Mexico achieved a surplus of \$5.5 billion instead of the deficit of \$3 billion planned under its IMF program. On the strength of a large trade surplus, Brazil will have reduced its current account deficit from \$14 billion in 1982 to less than \$2 billion in 1984. And although cutbacks of imports have played an important role in these turn-arounds, rising exports have also been important. For 8 major Latin American debtors, after declining by 8 percent from 1981 to 1983, export earnings rose by 12 percent in 1984. (Brazil's exports rose by 21 percent, and Mexico's non-oil exports by 35 percent; Morgan Guaranty, 1984.)

The centerpiece of debt recovery has materialized: recovery in the international economy. OECD growth has risen from  $-0.3$  in 1982 to 2.3 percent in 1983, and nearly 5 percent in 1984. There have been adverse developments as well: interest rates rose by 2 percentage points in early 1984 (before moderating), and the dollar has continued to climb instead of depreciating. Nonetheless, the benefits from higher OECD growth have swamped these negative factors. Because 1 percent extra OECD growth offsets 3 percent additional interest charges for non-oil developing countries (in the first year, and more if sustained over several years—although the relationship is less favorable for more highly indebted countries), higher growth in 1984 would have more than compensated for even a sustained 2 percent rise in LIBOR for the year. The most serious setback has been in debtor country growth: Latin American income declined 3 percent in 1983 and per-capita income is 10 percent below the 1980 level. However, growth has once again begun in the region in 1984, with rates of 2 to 3 percent in the major debtor countries, and the prospects are for higher domestic growth in the future.

Projections using updated information through mid-1984 reconfirm the broad analysis that the debt problem is manageable (see

my 1984 book, ch. 8). Even assuming some slowdown in OECD growth (2.7 percent in 1985, 2.3 percent in 1986, and 3 percent in 1987), relatively high interest rates (LIBOR at 12.5 percent in 1984, declining to 9.5 percent only by 1987); and assuming a 20 percent decline of the dollar phased over 1985–86 and oil at \$29 per barrel, the ratio of net debt to exports of goods and services declines from 200 percent in 1983 to 140 percent by 1987 for the 19 largest debtor countries. For Mexico the decline is from 310 percent to 210 percent; for Brazil, from 370 percent to 230 percent; and for Argentina, from a high 490 percent to 320 percent. Moreover, these projections provide for debtor-country growth at politically acceptable rates of  $4\frac{1}{2}$  percent or more per year. And the projections indicate expansion of nominal export earnings at a faster average rate (14.2 percent, 1984–87) than the level of the interest rate (about 12.5 percent including spread), meeting this important test (even though, with trade surpluses, export growth could be lower without increasing the debt-export ratio).

At the end of 1984, it appears that the debtor countries whose debt has systemic consequences are on track for recovery from the debt crisis.<sup>2</sup> External performance has been extremely strong for Brazil and Mexico. Argentina, a source of great concern earlier in the year, has reached agreement with the IMF. Mexico and Venezuela have obtained multiyear rescheduling from the banks; Brazil is likely to do so soon; and Argentina has reached tentative agreement with the banks for a shorter period. Nonetheless, for a secondary tier of some smaller countries, external sector pressures are more serious, in part because their exports are concentrated in commodities with weak prices. Chile and Peru face severe difficulties, although these should moderate once dollar depreciation and/or interest rate reductions permit more

normal recovery of metals prices. Bolivia, Nicaragua, Poland, Sudan, and Zaire—all on the bank regulators' list of "value impaired"—appear to have more protracted problems. Their aggregate debt is too small to threaten the system, however, and in some cases the best option may simply be to allow their arrears to accumulate.

Other studies have also reached the conclusion that broad recovery from the debt crisis may be expected (Morgan Guaranty; International Monetary Fund, 1984; Ronald Leven and David Roberts, 1983). Some critics have challenged the responsiveness of debtor country exports to OECD growth (Rudiger Dornbusch, 1984), but most studies tend to confirm an average volume elasticity in the range of 2 (the value in my model) and sufficient terms of trade response to boost the (average) revenue elasticity to approximately 3 (see the studies just cited and Carlos Diaz-Alejandro, 1984).<sup>3</sup>

Some critics have challenged such projections on grounds that the chances are small that the international economic variables will turn out exactly as assumed. However, this criticism loses force if one considers that if departures from projected values of exogenous variables are randomly distributed, favorable departures in some (such as OECD growth in 1984) will offset adverse departures for others (such as interest rates and dollar strength in 1984).

Some important conceptual underpinnings of this outlook require clarification. First, it assumes interest payments will exceed new borrowing (there will be an "outward transfer of resources"). While some question the political viability of this assumption, this outcome is prudent and necessary to reduce debt-export ratios from levels associated with overborrowing in the past. Moreover, with

<sup>2</sup> Brazil (\$92 billion total debt); Mexico (\$88 billion); Argentina (\$44 billion); and Venezuela (\$34 billion) account for three-fourths of bank debt owed by countries with debt-servicing disruptions in 1982–84 (see my 1984 study, pp. 25; 165).

<sup>3</sup> Thomas Enders and Richard Mattione (1984) project improvement on debt, but predict, somewhat questionably, painfully low Latin American domestic growth. The Inter-American Development Bank (1984) projects massive deficits even if Latin American growth is merely sufficient to keep up with population; but the study's parameters and assumptions stand in doubt (see my 1984 study, pp. 172–74).

domestic growth led by the export sector, an interim period of outward resource transfer need not mean slow domestic growth.

Second, the projections involve lower real import-GDP ratios than in the early 1980's for some major countries. Argentina, Mexico, and Venezuela had bloated import levels now corrected by real depreciation, and even allowing for significant rebounds, imports should not return to earlier levels. Brazil's sharp increase in production of oil and other import substitutes has accomplished the same objective. Adjustment is shifting to the expenditure-switching phase (shift in resource allocation from nontradables to exports and import substitutes), so that domestic growth can be higher than in the earlier expenditure-reducing (recessionary) phase.

Third, a caveat is that domestic inflation may have already become a more binding constraint on growth than external debt. Brazil's inflation exceeds 200 percent and Argentina's 700, and monetary-fiscal restraint to reduce inflation may hold back near-term growth, while politically debt may be blamed.

#### IV. Policy Implications

Despite favorable emerging economic reality, the debt problem remains politically vulnerable. The lagged effects of severe recession in 1983, and the runup in interest rates early in 1984, have caused political pressure, as shown by the emergence of the Cartagena group of Latin American debtor countries. The best strategy for the international financial system would be to pursue additional measures, to consolidate the gains made to date and insure against political disruption from possible future setbacks.

Macroeconomic policy in the North must ensure continued recovery. The U.S. fiscal deficits should be cut to permit more relaxed monetary policy, lower interest rates, and a lower dollar; and Europe and Japan should avoid further fiscal tightening until recovery is consolidated. A critical threshold of approximately  $2\frac{1}{2}$  percent for OECD growth is still the key to managing international debt (although the strong performance of 1984 leaves room for modest deceleration).

Sweeping measures on debt should be avoided. Write-downs and consolidation of debt in a new international agency would not only require scarce public capital and injure the debtors' credit ratings, but would also choke off the incentive for involuntary bank lending, as banks no longer would hold claims on the country. Interest forgiveness (such as a cut of all interest rates in half) would do great damage to the (highly leveraged) banks without providing much growth benefit for debtor countries (especially after deducting the loss of new lending).

For their part, banks need to continue aggregate new lending to developing countries at perhaps \$20 billion annually. They should extend "Mexico-type" (September, 1984) packages of reduced spreads above LIBOR and multiyear reschedulings, especially to countries that demonstrate effective adjustment. In addition, the banks would do well to provide some insurance against interest rate surges by offering a "Reimbursable Interest-Averaging Cap," whereby interest above a given ceiling (set near the initial market rate) would be deferred if international interest rates rose, and repaid once rates declined again. Banks could charge some spread premium for this benefit, and even obtain private insurance of amounts deferred.

The International Monetary Fund could usefully introduce a Compensatory Finance Facility for interest rate surges, to complement its facility for export fluctuation. Expanded resources are necessary for lending by the World Bank, regional development banks, and export credit agencies, so that the official sector can help redress the excessive shift from official to private financial sources in the past several years. Finally, industrial countries must avoid increased protection against imports from developing countries.

In sum, the major debtor countries have made much progress in recovering from the debt crisis. International management of the crisis on a case-by-case basis has amounted to a coherent strategy that has in fact shown favorable results so far. Nonetheless, considerable risk remains. The most serious, but not necessarily most likely, risk is from a collapse in OECD growth, the motor force in

debt recovery. The sharp slowdown in U.S. economy in late 1984 is troubling. Nonetheless, the likely absence of further oil shocks in the 1980's, and the fact that growth now begins from a base of much-reduced inflation, suggest that international growth rates closer to those achieved in the 1950's and 1960's than those of 1974-82 should be attainable. Overall, a series of moderate measures along the lines suggested here should be sufficient to consolidate the political and economic basis for continued recovery from the debt crisis, making it possible for international lending to return to a much more normal basis by the late 1980's.

#### REFERENCES

- Cline, William R. "International Debt and the Stability of the World Economy," in *Policy Analyses in International Economics*, No. 4, Institute for International Economics, September 1983.
- \_\_\_\_\_, *International Debt: Systemic Risk and Policy Response*, Washington: Institute for International Economics, 1984.
- Diaz-Alejandro, Carlos, "In Toto, I Don't Think We are in Kansas Anymore," prepared for Brookings Panel on Economic Activity, September 13-14, 1984.
- Dornbusch, Rudiger, "The Effects of OECD Macroeconomic Policies on Non-Oil LDCs: A Review," mimeo., MIT, 1984.
- Enders, Thomas O. and Mattione, Richard P., *Latin America: The Crisis of Debt and Growth*, Washington: The Brookings Institution, 1984.
- Leven, Ronald, and Roberts, David L., "Latin America's Prospects for Recovery," *Quarterly Review*, Federal Reserve Bank of New York, Autumn 1983, 8, 6-13.
- Federal Financial Institutions Examination Council, "Country Exposure Lending Survey: June 1984," Washington, 1984.
- Inter-American Development Bank, *External Debt and Economic Development in Latin America*, Washington, 1984.
- International Monetary Fund, *World Economic Outlook: 1984*, Washington, 1984.
- Morgan Guaranty, *World Financial Markets*, October/November 1984.

# Latin American Debt: Lessons and Pending Issues

By EDUARDO WIESNER\*

Two years have gone by since the international debt crisis surfaced in late 1982. Although this brief period may not provide enough perspective to draw firm conclusions, it would seem sufficient for arriving at some understanding of the causes and consequences of the crisis. In this paper I shall try to identify some of the lessons that can be derived from this experience. I will also look into the direction in which the current situation is evolving to see what are some of the main emerging issues. Needless to say, the paper does not pretend to be exhaustive or conclusive. Its principal purpose is to contribute to that continuing exercise whereby theory is tested by its predictive capabilities. The particular experience of the debt crisis seems to offer fertile ground to engage in this exercise because, as Jeffrey Sachs has stated, "it points up a lesson relearned several times in the past one hundred and fifty years: the international loan markets function differently from the textbook model of competitive lending" (1984, p. 1).

The paper is organized as follows. After succinctly reviewing the factors behind the debt crisis, I submit that the traditional absorption approach should be complemented to take into account the difficulties of developing countries in absorbing external savings efficiently. Then, I discuss the role of risk in the debt problem. It is in these two particular areas where I believe some of the most significant lessons can be extracted. Finally, I comment briefly on some of the issues that currently confront policymakers in Latin American countries.

## I. The Fiscal Origin of the Debt Crisis

No other set of factors explains more of the debt crisis than the fiscal deficits incurred by most of the major countries in Latin America. Although there were other factors which were relevant, I have no doubt that the main problem was excessive public (and private) spending that was financed by both easy domestic credit policies and by ample resources from abroad. The world recession and high real rates of interest in international markets aggravated the crisis, but I do not believe they created it. What actually happened was that previous domestic macroeconomic policies had made economies more vulnerable to exogenous factors.

The figures for the period 1979-82 cannot be more eloquent. In only four years, the three largest countries of the region—Argentina, Brazil, and Mexico—more than doubled the size of their nonfinancial public sector deficits, which rose from the already high levels of around 6 percent of *GDP* to well over 15 percent. Behind these growing fiscal deficits were strong political pressures for higher public spending. As long as external financing permitted total absorption to exceed domestic income, it was possible to accommodate those demands. But as the world recession worsened, and as it became evident that the additional financing from abroad was not being accompanied by a corresponding increase in exports or in domestic capital formation, capital inflows dropped substantially and the fiscal imbalance became an exchange rate and debt crisis. In brief, the debt crisis can be traced back to a fiscal disequilibrium, and ultimately to an unresolved political struggle between competing groups that wanted to have a larger share of income.

This view of the debt crisis as a fiscal manifestation of a political struggle may explain why it is still so difficult to resolve some of the most severe problems of the

\*Director, Western Hemisphere Department, Room 10-100, International Monetary Fund, Washington, D.C. 20431. The opinions expressed here are my own and do not necessarily reflect the views of the International Monetary Fund.



adjustment process. The intractability would arise because at the bottom of the adjustment process there still lies partially unsettled the question of who is going to pay for that portion of absorption which was not covered by taxes, has not become a productive investment within the debtor country, was presumably and to some extent a creditor risk, and may be by now gone forever. Under peremptory, abrupt adjustment, these political questions are resolved without delay by force of circumstances. But under managed adjustment, which is clearly preferable, the attempt to control the process may also allow, in some cases, for postponement of at least part of the most difficult political decisions, and thus it runs the risk of prolonging uncertainty and making recovery more elusive. This is a delicate political issue. Its importance lies in that it may explain why sometimes the gradual approach to adjustment does not seem to work.

## II. The Absorption of Foreign Financing

The combined external debt of Argentina, Brazil, Chile, Mexico, Peru, and Venezuela more than doubled between 1978 and 1982. This huge inflow of external resources was, of course, largely the other face of the large fiscal and current account deficits. External financing was allowing these countries to sustain absorption levels higher than national income. That is, domestic consumption and investment exceeded national income. Since most of the external financing came from private international banks, it was widely believed then that these countries were creditworthy and were effectively utilizing external savings. In some instances, the external financing was even allowing (if not inducing) real appreciation of the exchange rate and was luring policymakers into believing that a new era of low inflation, high growth, and exchange rate stability was about to begin. B. Kavalsky and L. Squire et al. (1984) point out that in countries like Peru, Argentina, and Mexico, real exchange rates actually appreciated. We now know that little of this proved to be a reality. Unfortunately, debt accumulation was not, according to Rudiger Dornbusch and Stanley

TABLE 1—SAVINGS AND INVESTMENT  
(As percent of *GDP*)

	1980	1981	1982	1983
<b>Argentina</b>				
Gross domestic investment	25.7	20.0	18.1	17.1
National savings	22.6	16.2	13.0	13.3
External savings	3.1	3.8	5.0	3.8
<b>Brazil</b>				
Gross domestic investment	21.1	19.2	18.4	15.0
National savings	16.0	15.1	12.9	12.0
External savings	5.1	4.1	5.5	3.0
<b>Mexico</b>				
Gross domestic investment	28.1	29.0	21.2	16.5
National savings	24.0	23.2	17.8	20.2
External savings	4.1	5.8	3.4	-3.7

Source: International Monetary Fund.

Fischer, "financing investment that was productive by the test of net dollar earnings" (1984, p. 4). After the initial appreciation of local currencies capital inflows turned into capital outflows (Arnold Harberger, 1984) precipitating sharp depreciations of the real exchange rate.

While the flows of external financing were pouring into the countries, two fundamental developments were taking place. The first, domestic savings as a proportion of *GDP*, was not increasing and, in most cases, was actually declining. Furthermore, gross domestic investment actually declined as a percentage of *GDP*. Countries were absorbing external resources (Ricardo Ffrench-Davis, 1983), but they were not using all of them to increase investment but rather to allow for more consumption. Since the accumulation of external debt was growing at a faster pace than the accumulation of productive investment,<sup>1</sup> a debt crisis was just a question of time, which would prove shorter in the event of external monetary disturbances. (See Table 1.)

The second very important development that was taking place had to do with the weakening, (B. A. de Vries, 1983) of the link

<sup>1</sup>H. J. Kharas holds that "the critical relationship in debt analysis is that between the accumulation of productive capital relative to external debt" (1984, p. 419).

between external financing and specific project investment evaluation. Since most of the additional external financing came from international commercial banks, and since the debtors generally were public sectors, the importance of the economic feasibility of investment projects was diluted. This meant that the global quality of the investment programs deteriorated and that many projects were initiated not because they were economically sound but because they had found external commercial financing which, directly or indirectly, was guaranteed by a sovereign debtor. And as Alexander Swoboda has pointed out, banks are "willing to acquire assets at rates of return lower than the riskiness of these assets" (1985, p. 24), if there is a guarantee. This development contrasted sharply with the traditional external financing for specific investment projects provided by multilateral agencies like the World Bank or the Inter-American Development Bank.

The lessons from this are not novel in my view, and yet recent experience would suggest that they need to be emphasized. To begin with, developing countries do not have an unlimited capacity to absorb external resources, because their absorptive investment capacity sets a limit to the volume of resources that they can utilize efficiently.<sup>2</sup> There are limits to the taxing capacity of developing countries' public sectors. Higher fiscal deficits may end up being financed by external savings, but this does not necessarily mean that all of these savings (Donough McDonald, 1982) become investments in the debtor countries or that the margin created by those additional external resources is used to increase investment.

### III. The Role of Risk Revisited

Although the debt crisis can be explained mainly in terms of fiscal deficits and inadequate economic policies in the debtor countries, it could also be explained by the wrong

perceptions that debtors and creditors had about the real risks they were taking. Had they had perfect foresight about these risks, no major crisis could have developed. A serious crisis can only result from a cumulative process. In the case of the current debt crisis, fiscal deficits and external debt accumulated because the risk factor did not play *ex ante*, in the early stages, its crisis-defusing role.

The experience of this debt crisis teaches that at least two of the fundamental risk assumptions that guided international lending and borrowing were wrong. First, international banks assumed that a sovereign risk was a small risk because public sectors normally do not default. As a result, they did not pay sufficient attention to the quality of economic policies in the debtor countries nor did they worry about the economic feasibility of the investment projects they were, directly or indirectly, financing. Apparently, they thought that an, implicit or explicit, official guarantee was all that was needed to circumvent risk. Perhaps it was considered that high spreads—internal and external—were sufficient to cover for any possible extra risk incurred. Finally, it is not clear what the risk perception of international banks was with regard to their lending to private banks or enterprises in the debtor countries.

Second, governments understood that if private flows were freely coming in to finance their expenditures, then their macroeconomic policies could not be all that bad and their exchange rates could not be much out of line with long-term equilibrium rates. In fact, they often found it difficult to control expenditure and to follow more prudent macroeconomic policies simply because the external financing was there. On the other hand, public sectors thought they were not taking any risks in the case of private external lending to their private sectors. Thus, they did not establish adequate controls or precise conditionality for that modality of international financing. The model<sup>3</sup> of international

<sup>2</sup>Sidney Alexander in his celebrated article did not imply that investment necessarily increased when external financing was available. His assertion was that both investment and consumption taken together could be higher than national income (1952, p. 265).

<sup>3</sup>See A. W. Corden for an interesting discussion on the distinction "between private problems and public problems" in the area of balance of payments (1977, p. 45).

lending held that these transactions could not pose a risk to public sectors, since, as Tim Congdon has put it, "there is no such thing as a balance of payments problem between consenting adults" (1982, p. 5).

We now have relearned that public sectors of borrowing countries can find themselves unable to service their debts and that, in some cases, debtor countries have been unable to resist pressures to guarantee private loans or to provide some form of exchange rate guarantee. This has converted a private risk into a fiscal burden. When I say that now we know the obvious, I am, of course, speaking within the context of what I have called above the investment absorption capacity of less developed countries. Nothing very original. It used to be almost an axiom a few years ago when few, if anyone, thought that developing countries could absorb massive doses of financial resources. They did not have an infinite amount of economically feasible investment projects, nor could they hope to have instant improvement in the quality of their macroeconomic policies. These realities, that meant risks, were temporarily forgotten.

Risk is a stabilizing factor in the functioning of markets. The insurance or guarantee of loans against risk may actually increase willingness to incur risk.<sup>4</sup> This is the so-called moral hazard problem. Attempts to interfere with the role of risk is a very risky exercise like the current crisis has confirmed. The process of delinking risk from the economic feasibility of investment projects and from the capacity of debtor countries to rapidly adopt all of the right macroeconomic policies in the face of abundant external financing is one of the most underestimated causes of the debt crisis. In brief, both debtors and creditors made wrong decisions on the basis of wrong perceptions of their respective risks.

One basic question emerges out of the above analysis. If both debtors and creditors share responsibility in the formation of the debt crisis, how are the costs of the resolution of that crisis going to be apportioned?

In principle, strict adherence to the role of risk would mean that the market would distribute the costs of resolving the crisis. In essence, this would mean that creditors would sell their claims in the market—and thus absorb in losses the reduced value of these claims—and divorce themselves from their debtors. For the debtors, the market solution would mean instant adjustment without financing and according to Jonathon Eaton and Mark Gersovitz, "an embargo on future loans" (1981, p. 290).

These hypothetical scenarios surely have been considered by both debtors and creditors since they would be the consequence of not providing any rescheduling or new money, on the one hand, and of defaulting on the other. The answer that both debtors and creditors have decided to give to the debt crisis is to manage jointly the resolution of their conflicting interests. The fact that both parties have chosen or resigned themselves to this route can be interpreted as an *ad hoc* market solution in which they have carefully assessed the risks of all alternatives and have come to the conclusion that their losses are minimized under this alternative.

#### IV. Choices and Issues

At this juncture, critical choices are being made in at least the following two main areas: international borrowing and lending; and the openness and efficiency of the economies of the Latin American countries. The first choice concerns the question of what to do about borrowing. While it is true, as William Cline has said, that "these countries have erred on the side of excessive borrowing in the last decade" (1984, p. 178), the question now is how to reduce the rate and, if possible, the level of indebtedness—which is the desirable thing to do—if at the same time there is a need to mitigate the tribulations of adjustment. This is a formidable dilemma. More financing alleviates adjustment, but simultaneously adds to the debt service. There is no simple solution to this predicament. At the bottom of the circularity involved there lies, once again, the question of how should the costs of the resolution of the crisis be apportioned. Until now, both parties have opted for negotiations rather

<sup>4</sup>See Henry Wallich for an excellent discussion of the risks and advantages of insurance on bank lending (1984).

than confrontation. While it is not possible to predict what will be the final outcome, it could be said that the answer will largely depend on what happens to the world economic recovery and to the export earnings of indebted countries. In the words of Mario Simonsen, "a growing economy with expanding exports hardly would seek confrontation with its creditors" (1984, p. 68).

The second area where difficult choices have to be made concern the degree of openness and of economic efficiency of the region. The issue here is how to preserve long-term efficiency when short-term and urgent pressures make it attractive to reduce openness and to seek economic reactivation through protectionism. The prospects in this area are not encouraging. Unfortunately, it seems that one of the most costly consequences of the debt crisis is that, at the end, some Latin American countries will be more closed and less efficient. At this juncture it seems that some policymakers of the region are striving to avoid a repetition of the crisis of 1982 and 1983. Apparently, they find it preferable to accumulate international reserves than to increase imports and to open up trade; they find it less destabilizing—and more fiscally productive—to raise import tariffs than to promote exports through the exchange rate. And in the future, they may feel less vulnerable if they regulate or control international capital flows than if they open up and let market risk, once again, be the final arbiter. Unfortunately, on the basis of the recent experience, some policymakers are apprehensive about what to expect from the free play of the market forces.

#### REFERENCES

- Alexander, Sidney, "Effects of a Devaluation on a Trade Balance," *IMF Staff Papers*, April 1952, 2, 263–78.
- Cline, William R., *International Debt: Systematic and Policy Response*, Washington: Institute for International Economics, 1984.
- Congdon, Tim, "A New Approach to the Balance of Payments," *Lloyd's Bank Review*, October 1982.
- Corden, A. W., *Inflation Exchange Rates and the World Economy*, Chicago: University of Chicago Press, 1977.
- de Vries, B. A., "International Ramifications of the External Debt Situation," *Amex Bank Review*, Special Papers, November 1983.
- Dornbusch, Rudiger and Fischer, Stanley, "The World Debt Problem," Report to the Group of Twenty-Four, UNDP/UNCTAD, September 1984.
- Eaton, Jonathan, and Gersovitz, Mark, "Debt with Potential Repudiation: Theoretical and Empirical Analysis," *Review of Economic Studies*, April 1981, 48, 289–309.
- Ffrench-Davis, Ricardo, "Que Paso con la Economia Chilena?" *Estudios Publicos*, Santiago, Chile, No. 11, 1983.
- Harberger, Arnold C., "Lessons for Debtor Country Managers and Policymakers," in J. T. Cuddington, and G. W. Smith, eds., *International Debt and the Developing Countries*, Washington: World Bank, 1985.
- Kavalsky, B. and Squire, L., et al., "Debt and Adjustment in Selected Developing Countries," unpublished paper, World Bank, July 1984.
- Kharas, H. J., "The Long-Run Creditworthiness of Developing Countries: Theory and Practice," *Quarterly Journal of Economics*, August 1984, 99, 415–40.
- McDonald, Donough, "Debt Capacity and Developing Country Borrowing: A Survey of the Literature," *IMF Staff Papers*, December 1982, 29, 603–46.
- Sachs, Jeffrey, "Theoretical Issues in International Borrowing," *Princeton Studies in International Finance*, No. 54, July 1984.
- Simonsen, Mario, "The Developing Country Debt Problem," unpublished paper, Getulio Vargas Foundation, June 1984.
- Swoboda, Alexander, "Debt and the Efficiency and Stability of the International Financial System," in J. T. Cuddington and G. W. Smith, eds., *International Debt and the Developing Countries*, Washington: World Bank, 1985 forthcoming.
- Wallich, Henry, "Insurance of Bank Lending to Developing Countries," Group of Thirty, Occasional Paper No. 15, New York, 1984.

## THE USE AND ABUSE OF ECONOMETRICS<sup>†</sup>

### Data and Econometricians—The Uneasy Alliance

By ZVI GRILICHES\*

Econometricians have an ambivalent attitude towards economic data. At one level, the "data" are the world that we want to explain, the basic facts that economists purport to elucidate. At the other level, they are the source of all our troubles. Their imperfection makes our job difficult and often impossible. Many a question remains unresolved because of "multicollinearity" or other sins of the data. We tend to forget that these imperfections give us our legitimacy in the first place. If the data were perfect, collected from well-designed randomized experiments, there would be hardly room for a separate field of econometrics. Given that it is the "badness" of the data that provides us with our living, perhaps it is not all that surprising that we have shown little interest in improving it, in getting involved in the grubby task of designing and collecting original data sets of our own. Most of our work is on "found" data, data that have been collected by somebody else, often for quite different purposes.

Economic data collection started primarily as a by-product of other governmental activities: tax and customs collections. Early on, interest was expressed in prices and levels of production of major commodities. Besides tax records, population counts, and price surveys, the earliest large-scale data collection efforts were various censuses, family expenditure surveys, and farm cost and production surveys. By the middle 1940's the overall economic data pattern was set:

governments were collecting various quantity and price series on a continuous basis, with the primary purpose of producing aggregate level indicators such as price indexes and national income accounts series, supplemented by periodic surveys of population numbers and production and expenditure patterns to be used primarily in updating the various aggregate series. Little micro data was published or accessible, except in some specific subareas, such as agricultural economics.

A pattern was also set in the way the data were collected and by whom they were analyzed. With a few notable exceptions, such as France and Norway, and until quite recently, econometricians were not to be found inside the various statistical agencies, and especially not in the sections that were responsible for data collection. Thus, there grew up a separation of roles and responsibilities. "They" collect the data and they are responsible for all their imperfections. "We" try to do the best with what we get, to find the grain of relevant information in all the chaff. Because of this, we lead a somewhat remote existence from the underlying facts we are trying to explain. We did not observe them directly; we did not design the measurement instruments; and, often we know little about what is really going on (for example, when we estimate a production function for the cement industry from census data without ever having been inside a cement plant).

In this we differ quite a bit from other sciences (including observational ones rather than experimental) such as archeology, astrophysics, biology, or even psychology where the "facts" tend to be recorded by the professionals themselves, or by others who have been trained and are supervised by those who will be doing the final data analysis.

<sup>†</sup>*Discussants:* Dale W. Jorgenson, Harvard University; A. H. Studenmund, Occidental College.

\*Professor of Economics, Harvard University, 125 Littauer Center, Cambridge, MA 02138. This paper is adapted from my larger manuscript (1985) which can be seen for more detail, examples, and references.

Economic data tend to be collected (or often more correctly "reported") by firms and persons who are not professional observers and who do not have any stake in the correctness and precision of the observations they report. While economists have increased their use of surveys in recent years and have even begun designing and commissioning special purpose surveys of their own, in general, the data collection and thus the responsibility for the quality of the collected material is still largely delegated to census bureaus, survey research centers, and similar institutions, and is divorced from the direct supervision and responsibility of the analyzing team.

It is only relatively recently, with the initiation of the negative income tax experiments and various longitudinal surveys intended to follow up the effects of different governmental programs, that econometric professionals have actually become involved in the primary data collection process. Once attempted, the job turned out to be much more difficult than was thought originally, and taught us some humility. Even with relatively large budgets, it was not easy to figure out how to ask the right question and to collect relevant answers. In part this is because the world is much more complicated than even some of our more elaborate models allowed for, and partly because economists tend to formulate their theories in non-testable terms, using variables for which it is very hard to find empirical counterparts. For example, even with a large budget, it is difficult to think of the right series of questions, answers to which would yield an unequivocal number for *the* level of "human capital" or "permanent income" of an individual. Thinking about such "alibi-removing" questions should make us a bit more humble, restrain our continuing attacks on the various official data-producing agencies, and push us towards formulating theories with more regard to what is observable and what kind of data may be available.

Even allowing for such reservations, there has been much progress over the years as a result of the enormous increase in the quantity of data available, in our ability to manipulate them, and in our understanding

of their limitations. Especially noteworthy have been the development of various longitudinal micro data sets (such as the Michigan PSID tapes, the Ohio State *NLS* surveys, the Wisconsin high school class follow-up study, and others), the computerization of the more standard data bases and their more easy accessibility at the micro, individual response level (I have in mind here such developments as the Public Use Samples from the *U.S. Census of Population* and the *Current Population Surveys*). Unfortunately, much more progress has been made with labor force and income data, where the samples are large, than in the availability of firm and other market transaction data.

While significant progress has been made in the collection of financial data and security prices, as exemplified in the development of the CRISP and Compustat data bases which have had a tremendous impact on the field of finance, we are still in our infancy as far as our ability to interrogate and get reasonable answers about other aspects of firm behavior. Most of the available micro data at the firm level are based on legally required responses to questions from various regulatory agencies who do not have our interests exactly in mind.

We do have now, however, a number of extensive longitudinal micro data sets that have opened a host of new possibilities for analysis and also raised a whole range of new issues and concerns. After a decade or more of studies that try to use such data, the results have been somewhat disappointing. We, as econometricians, have learned a great deal from these efforts and developed whole new subfields of expertise, such as sample selection bias and panel data analysis. We know much more about these kinds of data and their limitations, but it is not clear that we know much more about the roots and modes of economic behavior that underlie them.

The encounters between econometricians and data are frustrating and ultimately unsatisfactory, both because econometricians want too much from the data and hence tend to be disappointed by the answers, and because the data are incomplete and imperfect. In part it is our fault, the appetite grows with

eating. As we get larger samples, we keep adding variables and expanding our models, until on the margin, we come back to the same insignificance levels.

There are at least three interrelated and overlapping causes of our difficulties: 1) the theory (model) is incomplete or incorrect; 2) the units are wrong, either at too high a level of aggregation or with no way of allowing for the heterogeneity of responses; and 3) the data are inaccurate on their own terms, incorrect relative to what they purport to measure. The average applied study has to struggle with all three possibilities.

At the macro level and even in the usual industry level study, it is common to assume away the underlying heterogeneity of the individual actors and analyze the data within the framework of the "representative" firm or "average" individual, ignoring the aggregation difficulties associated with such concepts. In analyzing micro data, it is much more difficult to evade this issue and hence much attention is paid to various individual "effects" and "heterogeneity" issues. This is where longitudinal data can be of help—in their ability to control and allow for additive individual effects. On the other hand, as is the case in most other aspects of economics, there is no such thing as a free lunch: going down to the individual level exacerbates both some of the left-out variables problems and the importance of errors in measurement. Variables such as age, land quality, or the occupational structure of an enterprise, are much less variable in the aggregate. Ignoring them at the micro level can be quite costly, however. Similarly, measurement errors which tend to cancel out when averaged over thousands or even millions of respondents, loom much larger when the individual is the unit of analysis.

It is possible, of course, to take an alternative view: that there are no data problems, only model problems in econometrics. For any set of data there is the "right" model. Much of econometrics is devoted to procedures which try to assess whether a particular model is right in this sense and to criteria for deciding when a particular model fits and is "correct enough." Theorists and model builders often proceed, however, on the as-

sumption that ideal data will be available and define variables that are unlikely to be observable, at least not in their pure form. Nor do they specify in adequate detail the connection between the actual numbers and their theoretical counterparts. Hence, when a contradiction arises, it is then possible to argue "so much worse for the facts."

In practice one cannot expect theories to be specified to the last detail nor the data to be perfect or of the same quality in different contexts. Thus any serious data analysis has to consider at least two data generation components: the economic behavior model describing the stimulus-response behavior of the economic actors and the measurement model describing how and when this behavior was recorded and summarized. While it is usual to focus our attention on the former, a complete analysis must consider them both. (At this point in the longer manuscript, there follows a review of types of economic data, notions of data "quality," and a discussion of the major problems associated with the use of different sources and types of data.)

Over thirty years ago, Oscar Morgenstern (1950) asked whether economic data were accurate enough for the purposes that economists and econometricians were using them for. He raised serious doubts about the quality of many economic series and implicitly about the basis for the whole econometrics enterprise. Years have passed and there has been very little coherent response to his criticisms.

There are basically four responses to his criticism and each has some merit. 1) The data are not that bad. 2) The data are lousy but it does not matter. 3) The data are bad but we have learned how to live with them and adjust for their foibles. 4) That is all there is—it is the only game in town and we have to make the best of it.

There clearly has been great progress both in the quality and quantity of the available economic data. In the United States, much of the agricultural statistical data collection has shifted from judgment surveys to probability-based survey sampling. The commodity coverage in the various official price indexes has been greatly expanded and much more attention is being paid to quality change

and other comparability issues. Decades of criticisms and scrutiny of official statistics have borne some fruit. Also, some of the aggregate statistics have now much more extensive micro data underpinnings. It is now routine in the United States to collect large periodic labor force activity and related topics surveys and release the basic micro data for detailed analysis with relative short lags. But both the improvements in and the expansion of our data bases have not really disposed of the questions raised by Morgenstern. As new data appear and as new data collection methods are developed, the question of accuracy persists. While quality of some of the "central" data has improved, it is easy to replicate some of Morgenstern's horror stories even today. For example, in 1982 the U.S. trade deficit with Canada was either \$12.8 or \$7.9 billion depending on whether this number came from U.S. or Canadian publications. It is also clear that the national income statistics for some of the LDCs are more political than economic documents.

Morgenstern did not distinguish adequately between levels and rates of change. Many large discrepancies represent definitional differences, and studies that are mostly interested in the movements in such series may be able to evade much of this problem. The tradition in econometrics of allowing for "constants" in most relationships and not overinterpreting them, allows implicitly for permanent "errors" in the levels of the various series. It is also the case that in much of economic analysis, one is after relatively crude first-order effects and these may be rather insensitive even to significant inaccuracies in the data. While this may be an adequate response with respect to much of the standard, especially macroeconomic analysis, it seems inadequate when we contemplate some of the more recent elaborate nonlinear multi-equational models being estimated at the frontier of the subject. They are much more likely to be sensitive to errors and inconsistencies in the data.

In the recent decade there has been a revival of interest in "error" models in econometrics, though the progress in sociology on this topic seems more impressive.

Recent studies using micro data from labor force surveys, negative-tax experiments, and similar data sources exhibit much more sensitivity to measurement error and sample selectivity problems. Even in the macro area there has been some progress, and the "rational expectations" wave has made researchers more aware of the discrepancy between observed data and the underlying forces that are presumably affecting behavior. All of this has yet to make a major dent on econometric textbooks and econometric teaching, but there are signs that change is coming. It is more visible in the areas of discrete variable analysis and sample selectivity issues than in the errors of measurement area per se, but the increased attention that is devoted to data provenance in these contexts is likely to spill over into a more general data "aware" attitude.

One of the reasons why Morgenstern's accusations were brushed off was that they came from "outside" and did not seem sensitive to the real difficulties of data collection and data generation. In most contexts the data are imperfect not by design, but because that is all there is. Empirical economists have over generations adopted the attitude that having bad data is better than having no data at all, that their task is to learn as much as is possible about how the world works from the unquestionably lousy data at hand. While it is useful to alert users to their various imperfections and pitfalls, the available economic statistics are our main window on economic behavior. In spite of the scratches and the persistent fogging, we cannot stop peering through it and trying to understand what is happening to us and to our environment, nor should we. The problematic quality of economic data presents a continuing challenge to econometricians. It should not cause us to despair, but we should not forget it either.

The availability of longitudinal micro data has helped by providing us with one way of controlling for missing but relatively constant information on individuals and firms. It is difficult, however, to shake off the impression that, here also, the progress of econometric theory and computing ability is outracing the increased availability of data



and our understanding and ability to model economic behavior in increasing detail. While we tend to look at the newly available data as adding degrees of freedom grist to our computer mills, the increased detail often raises more questions than it answers. Particularly striking is the great variety of responses and differences in behavior across firms and individuals. Specifying additional distributions of unseen parameters rarely adds substance to the analysis.

What is needed is a better understanding of the behavior of individuals, better theories and more and different variables. Unfortunately, standard economic theory deals with "representative" individuals and "big" questions, and does not provide much help in explaining the production or hiring behavior of a particular plant at a particular time, at least not with the help of the available variables. Given that our theories, while couched in micro language, are not truly micro oriented, perhaps we should not be

asking such questions. Then what are we doing with micro data? We should be using the newly available data sets to help us find out what is actually going on in the economy and in the sectors that we are analyzing, without trying to force our puny models on them. The real challenge is to try to stay open, to learn from the data, but also, at the same time, not drown in the individual detail. We have to keep looking for the forest among all these trees.

#### REFERENCES

- Griliches, Z., "Data Problems in Econometrics," NBER Technical Working Paper No. 39, July 1984; reprinted in his and M. Intriligator, eds., *Handbook of Econometrics*, Vol. III, Amsterdam: North-Holland, 1985 forthcoming, ch. 25.
- Morgenstern, Oscar, *On the Accuracy of Economic Observations*, Princeton: Princeton University Press, 1950.

# The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests

By DONALD N. MCCLOSKEY\*

Roughly three-quarters of the contributors to the *American Economic Review* misuse the test of significance. They use it to persuade themselves that a variable is important. But the test can only affirm a likelihood of excessive skepticism in the face of errors arising from too small a sample. The test does *not* tell the economist whether a fitted coefficient is large or small in an economically significant sense.

The criticism is distinct from the criticism that in a world of publication-counting deans, there is an incentive to mine the data, giggling uncomfortably when caught. Economists, being professional cynics, are much amused by data mining and significance fishing (Gordon Tullock, 1959; Edward Ames and Stanley Reiter, 1961; Edgar Feige, 1975; Edward Leamer, 1978; Thomas Mayer, 1980; Michael Lovell, 1983; Frank Denton, 1985). But the point here is that even under classical conditions the *t*-test is irrelevant much of the time.

Neither criticism is controversial or arcane. Statisticians, psychometricians, sociometricians, econometricians, and other metrical folk have understood them both for 60 years (see Kenneth Arrow, 1960; Zvi Griliches, 1976). Both should be parts of the statistical education of an economist, yet almost none of the texts in econometrics mention them.

## I. An Example: Tests of Purchasing Power Parity

The usual test of purchasing power parity (see J. R. Zecher and myself, 1984) fits prices

at home ( $P$ ) to prices abroad ( $P^*$ ), allowing for the exchange rate ( $e$ ):  $P = \alpha + \beta(eP^*) + \text{error term}$ . The equation can be in levels or rates of change. If the coefficient  $b$  is statistically significantly different from 1.0, the hypothesis of purchasing power parity is rejected; if not, not. The test seems to tell about substantive significance without any tiresome inquiry into how true a hypothesis must be in order to be true. The table of  $t$  will tell.

But a number is large or small relative only to some standard. Forty degrees of frost is paralyzing cold by the standard of Virginia, a normal day by the standard of Saskatoon in January, and a heat wave by the standard of most interstellar gas. A *New Yorker* magazine cartoon shows water faucets labeled "Hot (A Relative Concept)" and "Cold (A Relative Concept)." Nothing is large-in-itself. It is large (or yellow, rich, cold, stable, well-integrated, selfish, free, rising, monopolistic) relative to something with which it can be interestingly compared. The remark "But how large is large?" is one of those seminar standbys, applying to any paper, like "Have you considered simultaneity bias?" or "Are there unexploited opportunities for entry?" It's usually a good question, inheriting some of its excellence from its father in thought, the mind-stunning "So What?" (and its Jewish mother: "So What Else is New?"). You say the coefficient is 0.85 with a standard error of 0.07? So?

The literature does not discuss how near the slope has to be to 1.0 to be able to say that purchasing power parity succeeds or fails. It does not answer how large is large. The only standard offered is statistical significance, that is, how surprising it would be to get the observed sample if the hypothesis of  $\beta = 1.0$  were in fact exactly true.

But "exactly" true is not relevant for most economic purposes. What is relevant is merely that  $\beta$  is in the neighborhood of 1.0,

\*Departments of Economics and of History, University of Iowa, Iowa City, IA 52242. I thank Leanne Swenson for research assistance, and seminar participants at Iowa, Nebraska, Yale, Chicago, and McMaster for comments. The work was supported by the John Simon Guggenheim Foundation, the Institute for Advanced Study, and the Program in Humanities, Science and Technology of the National Endowment for the Humanities.

where "the neighborhood" is defined by *why* it is relevant—for policy, for academic reputation, for the progress of knowledge. The question requires thought about the loss function. One begins to think that the neighborhood of small loss might be large. And even outside it, one begins to think that  $\beta = .10$ , say, would still be economically significant, were the fit tight enough to constrain prices at home; or that even a coefficient of  $-7854.86$  would belie closed economy models of inflation.

The usual test does not discuss standards. It gives them up in favor of irrelevant talk about the probability of a type I error in view of the logic of random sampling. Most economists appear to have forgotten how narrow is the question that a statistical test of significance answers. It tells the intrepid investigator how likely it is that, *because of the small size of the sample he has*, he will make a mistake of excessive skepticism in rejecting a true hypothesis (in this case,  $\beta = 1.0$ ). Though not to be scorned, it isn't much. It warns him about a certain narrow kind of foolishness.

The elementary but neglected point is that statistical tests of significance are merely about one sort of unbiased errors in *sampling*. The standard error, after all, is  $(s^2/N)^{1/2}$ . Except in the limiting case of literally zero correlation, if the sample were large enough all the coefficients would be significantly different from everything. The inverse of the square root of an extremely large number is very small. Any social scientist with large samples has had such logic impressed on him by events. A psychologist, Paul Meehl, for instance, reports a sample of 55,000 Minnesota high school seniors which "reveal statistically significant relationships in 91 percent of pairwise associations among a congeries of 45 miscellaneous variables such as sex, birth order, religious preference, . . . , dancing, interest in woodworking. . . . The majority of variables exhibited significant relationships with *all but three of the others*, often at a very high confidence level" (1970, p. 259).

The large-sample case makes clear the irrelevance of statistical significance to the main question: So what? In the usual test of

purchasing power parity, a sample size of a million yielding a very tight estimate that  $\beta = 0.999$ , "significantly" different from 1.0, could be produced under the usual procedures as evidence that the theory had "failed." Common sense, presumably, would rescue the investigator from asserting that if  $\beta = 0.999$ , with a standard error of .00000001, we should abandon purchasing power parity, or run our models of the American economy without the world price level. Similar common sense should be applied to findings that  $\beta = .80$  or 1.30 with sample sizes of 30. It is not.

The point can be put most sharply by supposing that we *knew* the coefficient to be, say, 0.85. Suppose God told us. God does not play dice with the universe, and His is no mere probabilistic assurance. Would the scientific task be finished? No, it would not. We would still need to decide, by some criterion of why it matters (a human, not a divine, concern), whether 0.85 is high enough to affirm the theory. No mechanical procedure can relieve us of this responsibility. Nor is it a decision that should be made privately, as a matter of "mere opinion." It is the most important scientific decision, and it should be made out in the open. The test of significance doesn't make it.

## II. A History of Consciousness

The overuse of statistical significance arises largely from its name. Surely, it insinuates, we serious scientists should be interested first of all in "significant" coefficients: the wise and good would not wish to waste time on trivialities. The appeal is part of the rhetoric of statistics (compare my book, 1985, ch. 2). The British inventors of statistics, as recipients of classical educations, were skillful in naming their ideas. As William Kruskal, a statistician of note, has argued:

Suppose that Sir R. A. Fisher—a master of public relations—had not taken over from ordinary English such evocative words as "sufficient," "efficient," and "consistent" and made them into precisely defined terms of statistical theory. He might, after all, have used

utterly dull terms for those properties of estimators, calling them characteristics A, B, and C.... Would his work have had the same smashing influence that it did? I think not, or at least not as rapidly. [1978, p. 98]

As the words spread to less sophisticated research workers, the task of undoing the rhetorical damage commenced. The earliest paper making the point of the present one was written in 1919 by, alarmingly, Edwin Boring. Attacks on the mechanical use of significance became commonplace early in statistical education. By 1939, for example, a *Statistical Dictionary of Terms and Symbols* of no great intellectual pretensions was putting the point utterly plainly: "A significant difference is not necessarily large, since, in large samples, even a small difference may prove to be a significant difference. Further, the existence of a significant difference may or may not be of practical significance" (A. K. Kurtz and H. A. Edgerton, 1939, article "Significant Difference"). M. G. Kendall and A. Stuart's *Advanced Theory of Statistics* explicitly recognized the mischief in the rhetoric, recommending the phrase "size of the test" in preference to "significance level" (1951, p. 163n); the sociometricians Denton Morrison and Ramon Henkel (whose book *The Significance Test Controversy*, 1970, is the best reading on the subject) suggest that "significance test" be replaced by the less portentous "sample error decision procedure" (p. 198).

In the 1930's, Jerzy Neyman and E. S. Pearson, and then more explicitly Abraham Wald, argued that actual statistical decisions should depend on substantive, not merely statistical, significance. As Wald wrote in 1939:

The question as to how the form of the weight [i.e., loss] function  $W(\theta, \omega)$  should be determined, is not a mathematical or statistical one. The statistician who wants to test certain hypotheses must first determine the relative importance of all possible errors,

which will entirely depend on the special purposes of his investigation. [p. 302]

Economists have largely ignored Wald's economical logic, with the result that few textbooks in econometrics mention that the goodness or badness of a hypothesis cannot be decided on merely statistical grounds.

### III. The Practice of Economists

It is not easy, then, to justify the use of probabilistic models to answer nonprobabilistic questions. One might retort that good economists do not make such mistakes. But they do, as may be seen from their best practice, in this *Review*. From the fifty full-length papers using regression analysis in the four regular issues of 1981, 1982, and 1983, I took a sample of ten for close scrutiny. Since the purpose is to criticize a socially accepted practice, not to embarrass individual writers, I shall withhold the names here (a larger version of the paper contains them, and a still larger one will examine all fifty).

Of the ten papers, only two do not admit experimenting with the regressions, sometimes with hundreds of different specifications. None propose to alter their levels of significance. Only two of the ten do not use a sign test in conjunction with a significance test: the variable has a statistically significant coefficient *and the right (or expected) sign*. Little statistical theory seems to lie behind the practice, although it seems sensible enough—a beginning, indeed, of looking beyond statistical significance to the size of the coefficient. One of the papers uses a sample of convenience so convenient that it looks like a universe, about which sampling theory can tell nothing: all counties in Alabama, Mississippi, North Carolina, and South Carolina. Four of the ten use true samples, such as the opinions of 6,000 Swedes on the current and expected rate of inflation. The only doubt here is the disproportion of effort in dealing with sampling errors when others are probably more serious. At  $N = 6000$  we can surely dismiss Student and attend to bias. As Leamer remarked recently, "when the sampling uncertainty... gets small

compared to the misspecification uncertainty... it is time to look for other forms of evidence, experiments, or nonexperiments" (1983, p. 33). The other five papers use time-series. One can only ask quietly and pass on: from what universe is a time-series a random sample, and if there is such a universe, is it one we wish to know about?

The most important question is whether the economists in the sample mix up statistical and substantive significance. Even on purely statistical grounds the news is not good: none of the papers mention the word "power," though all mention "significance." Statisticians routinely advise examining the power function, but economists do not follow the advice. Some follow its spirit, avoiding the excessive gullibility of the type II error by treating the machinery of hypothesis testing with a certain reserve. Most do not. Only three of the ten do not jump with abandon from statistical to substantive significance. The very language, though mostly formulaic, sometimes exposes the underlying attitude. One paper slipped into using the phrase "statistically important."

Seven of the papers, then, let statistical significance do the work of substantive significance. Usually this is accomplished by a fallacy of equivocation. The result on page 10 that is (statistically) significant turns up as (economically) significant on p. 20. In the worst cases there is no attempt to show how large the effects are, or whether the statistical tests of their largeness are powerful, or what standard of largeness one should use. In four of the seven papers with significant errors in the use of significance there is some discussion of how large a coefficient would need to be to be large, but even these let statistical significance do most of the work. And even in the three papers that recognize the distinction and apply it consistently, there is flirtation with intellectual disaster. The siren song of "significance" is a hazard to navigation.

#### IV. What is to be Done?

If we do not wish to leave science to chance, we must rethink the use of statistical significance in economics. Econometrics

courses should teach the relevant decision theory, as judging from results they appear not now to do. It would help if the standard statistical programs did not generate *t*-statistics in such profusion. The programs might be written to ask "Do you really have a probability sample?," "Have you considered power?," and, above all, "By what standard would you judge a fitted coefficient large or small?" Or perhaps they could merely say, printed in bold capitals beside each equation, "So What Else is New?"

#### REFERENCES

- Ames, Edward and Reiter, Stanley, "Distributions of Correlation Coefficients in Economic Time Series," *Journal of the American Statistical Association*, September 1961, 56, 627-56.
- Arrow, Kenneth, "Decision Theory and the Choice of a Level of Significance for the *t*-Test," in Ingram Olkin et al., eds., *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, Stanford: Stanford University Press, 1960, 70-78.
- Boring, Edwin G., "Mathematical Versus Scientific Significance," *Psychological Bulletin*, 1919, 15, 335-38.
- Denton, Frank T., "Data Mining as an Industry," *Review of Economics and Statistics*, 1985 forthcoming.
- Feige, Edgar, "The Consequences of Journal Editorial Policies and a Suggestion for Revision," *Journal of Political Economy*, December 1975, 83, 1291-96.
- Griliches, Zvi, "Automobile Prices Revisited: Extensions of the Hedonic Hypothesis," in N. E. Terleckyj, ed., *Household Production and Consumption*, NBER *Studies in Income and Wealth*, Vol. 40, 1976, 325-90.
- Kendall, M. G. and Stuart, A., *Advanced Theory of Statistics*, 3rd ed., Vol. II, London: Griffin, 1951.
- Kruskal, William H., "Formulas, Numbers, Words: Statistics in Prose," *The American Scholar*, 1978; reprinted in D. Fiske, ed., *New Directions for Methodology in Social and Behavioral Sciences*, San Francisco: Jossey-Bass, 1981.

- Kurtz, A. K. and Edgerton, H. A., *Statistical Dictionary of Terms and Symbols*, New York: Wiley & Sons, 1939.
- Leamer, Edward, *Specification Searches: Ad Hoc Inferences with Nonexperimental Data*, New York: Wiley & Sons, 1978.
- , "Let's Take the Con Out of Econometrics," *American Economic Review*, March 1983, 73, 31–43.
- Lovell, Michael C., "Data Mining," *Review of Economics and Statistics*, February 1983, 45, 1–12.
- McCloskey, Donald N., *The Rhetoric of Economics*, Madison: University of Wisconsin Press, 1985.
- Mayer, Thomas, "Economics as a Hard Science: Realistic Goal or Wishful Thinking?," *Economic Inquiry*, April 1980, 18, 165–78.
- Meehl, Paul E., "Theory Testing in Psychology and Physics: A Methodological Paradox," *Philosophy of Science*, June 1967, 34, 103–15; reprinted in Morrison and Henkel, 1970.
- Morrison, Denton E. and Henkel, Ramon E. "Significance Tests Reconsidered," *American Sociologist*, May 1969, 4, 131–40, reprinted in Morrison and Henkel, 1970.
- and ———, *The Significance Test Controversy—A Reader*, Chicago: Aldine, 1970.
- Tullock, Gordon, "Publication Decisions and Tests of Significance: A Comment," *Journal of the American Statistical Association*, September 1959, 54, 593.
- Wald, Abraham, "Contributions to the Theory of Statistical Estimation and Testing Hypotheses," *Annals of Mathematical Statistics*, December 1939, 10, 299–326.
- Zecher, J. R. and McCloskey D. N., "The Success of Purchasing Power Parity," in M. D. Bordo and A. J. Schwartz, eds., *Retrospective on the Classical Gold Standard*, Chicago: University of Chicago Press, 1984.

# Macroeconometric Modeling and the Theory of the Representative Agent

By JOHN GEWEKE\*

The Lucas critique of econometric policy evaluation (1976) argues that, to the extent econometric models do not capture the primitive parameters of tastes and technology, their coefficients can be expected to vary with changes in policy regimes. Several econometricians have undertaken empirical work that separates the parameters of tastes, technology, and policy, estimating models that are in principle immune from this critique. Exemplary contributors to this effort have been Thomas Sargent (1978) and Sargent and Lars Hansen (1980). The empirical work inspired by the Lucas critique has proceeded using representative agent models and aggregate data. The treatment of expectations and dynamic optimization has been careful, although at times necessarily limited by the analytical requirements of attaining closed-form solutions for dynamic programming problems under uncertainty. Potential problems due to aggregation have usually been ignored, although when preferences and technology are quadratic it appears that they can be disposed of quickly: "assuming a representative firm is only a convenience, as the model admits a tidy theory of aggregation" (Sargent, 1978, p. 1016).

It is ironic that a paradigm that emphasizes the isolation of the primitive parameters of tastes and technology has led to empirical work that has been conducted almost exclusively with aggregate data. (There are exceptions: T. MaCurdy, 1983; J. Biddle, 1984.) There are several difficulties with this development. First, the fact that representative agent models with exact aggregation *can* be constructed is unrelated to whether or

not these models are adequate; we can also construct models in which agents' behavior is unaffected by the policy regime. To exclusively model and test one but not the other appears to be a misplacement of emphasis. Second, some of this work has proceeded using representative agents whose behavior cannot be aggregated exactly (see, M. Eichenbaum, Hansen, and Kenneth Singleton, 1984, for an example).

There is a third and most fundamental objection to empirical work which seeks to avoid the Lucas critique, yet uses aggregate data. Whenever econometric policy evaluation is undertaken using models estimated with aggregate data, it is implicitly presumed that the aggregator function is structural with respect to the policy intervention in question. Formally, aggregator functions are no more structural than are within-regime, reduced-form relations of endogenous to policy variables. As a modeling strategy, ignoring the sensitivity of aggregators to policy changes seems no more compelling than ignoring the dependence of expectations on the policy regime.

This paper describes a very modest econometric model in which the effects of ignoring aggregation and of ignoring expectations, each within the context of several representative agent models, can be appreciated. Objective functions are quadratic and prices are disparate across agents; in these respects the model's assumptions are in the mainstream of Lucas and subsequent empirical work. The model is carefully contrived so that exact aggregation is always possible. This is not essential to the argument in any way, but it drastically reduces the number of circumstances to be examined in constructing numerical examples. The example pertains to neoclassical production, and in each case there are three distinct representative agents: one for production, one for factor demand,

\*Professor of Economics, Duke University, Durham, NC 27706. Financial support from NSF Grant SES-8318778 and a Sloan Research Fellowship are gratefully acknowledged. A more detailed version of this paper is available on request from the author.

and one for supply. The three are not the same, and except for one easily understood exceptional case, all three give different and incorrect evaluations of the effect (here, on output) of a policy change (here, a subsidy). Numerical examples suggest that the aggregation and expectations problems are of about the same order of magnitude.

### I. Representative Agents and Aggregation over Prices

A small industry in a small country produces a single, homogeneous output  $y$  ultimately sold competitively in a world market. Output is produced using a single variable factor  $x$  that is purchased competitively in the domestic market. The production technology is the same in all firms in the industry: for the  $i$ th firm at time  $t$ ,  $y_{it} = ax_{it} + dx_{it}^2$ ;  $a > 0$ ,  $d < 0$ . Firms are geographically distributed throughout the small country, and because of this the price for the output of the  $i$ th firm is  $p_{it}$ , the price varying, say, with access to deep-water ports. Output price changes through time but relative output prices across firms never change,  $p_{it} = p_t p_i$ , as would be the case if price differentials were due to different transportation costs and the relative prices of output and transportation were unchanging. To normalize take  $E(p_i) = 1$  and assume  $E(p_i^j) = m_j$  exists for  $j = -3, -2, \dots, 2$ . Input price  $r_t$  is the same for all firms. Firms maximize current-period profits, so the  $i$ th firm's factor demand is  $x_{it} = -a/2d + (1/2d)r_t/p_{it}$  and its supply function is  $y_{it} = -a^2/4d + (1/4d)r_t^2/p_{it}^2$ .

An econometrician has available the aggregate time-series  $p_t$ , which is the average price received;  $y_t$ , the average output of a firm, and  $x_t$ , the average factor input of a firm, for  $t = 1, \dots, T$ . The aggregate supply function is  $y_t = -a^2/4d + (m_{-2}/4d)r_t^2/p_t^2$ , the aggregate factor-demand function is  $x_t = -a/2d + (m_{-1}/2d)r_t/p_t$ , and the aggregate production function is

$$y_t = -\left(\frac{a^2}{4d}\right)\left[m_{-2}/m_{-1}^2 - 1\right] \\ + \left(m_{-2}/m_{-1}^2\right)ax_t + \left(m_{-2}/m_{-1}^2\right)dx_t^2.$$

The seriousness of any misspecification can be gauged only in relation to the purposes for which the model is used. To that end, suppose the econometrician's model estimated with aggregate time-series is used to evaluate the effect of a subsidy on output. Consider two subsidy schemes: an ad valorem policy in which the firm faces the price  $sp_{it}$ ,  $s > 1$ , and an ad rem policy in which the price becomes  $p_{it} + u$ ,  $u > 0$ . Actual output with the ad valorem subsidy will be indicated correctly by the aggregate supply function with  $sp_t$  used in lieu of  $p_t$ :  $y_t = -a^2/4d + (m_{-2}/4d)r_t^2/(sp_t)^2$ . Hence the actual effect of an ad valorem subsidy  $s$  on output, when factor price is  $r_t$  and aggregate output price is  $p_t$ , is  $(s^{-2} - 1)(m_{-2}/4d)(r_t^2/p_t^2)$ . Since the aggregate supply function is exact and the subsidy represents no change in the distribution of output prices across firms that previously prevailed, this is unsurprising. If the effects of the subsidy are inferred from the representative agent for the aggregate factor-demand function then the anticipated effect is  $(s^{-2} - 1)m_{-1}r_t^2/(4dp_t^2)$ , and the ratio of the anticipated to the actual effect is  $m_{-1}/m_{-2} < 1$ . If the effects of the subsidy are inferred from the representative agent for the aggregate production function, then the ratio of the anticipated to the actual effect is  $m_{-1}^2/m_{-2}^2$ .

None of the three representative agents provides a correct evaluation of the effect of an ad rem subsidy. Actual output with ad rem subsidy  $u$  is  $y_t = -a^2/4d + (r_t^2/4dp_t^2)E[p_i + (u/p_i)]^{-2}$ . Hence the effect of the subsidy  $u$  on aggregate output  $y$  in period  $t$  is  $(r_t^2/4dp_t^2)E\{[p_i + (u/p_i)]^{-2} - p_i^{-2}\}$ . The anticipated effects, using our three representative agents, are  $(m_{-2}r_t^2/4dp_t^2)\{[1 + (u/p_t)]^{-2} - 1\}$  from the aggregate supply function;  $(m_{-1}r_t^2/4dp_t^2)\{[1 + (u/p_t)]^{-2} - 1\}$  from the aggregate factor-demand function; and  $(m_{-1}^2m_{-2}^2r_t^2/4dp_t^2)\{[1 + (u/p_t)]^{-2} - 1\}$  from the aggregate production function. The ratios of the anticipated effects to one another are the same as for the ad valorem subsidy. Standard Taylor-series methods show that the ratios of anticipated effects to true ones are  $m_{-2}/m_{-3}$ ,  $m_{-1}/m_{-3}$ , and  $m_{-1}^2/m_{-2}m_{-3}$ , respectively, for small  $u$ ; all are less than one.



## II. Representative Agents and Price Expectations

Problems of a similar kind arise when firms plan production based on anticipated rather than actual prices, but the econometrician ignores this fact. Suppose that firms now receive the same price  $p_t$  for their output, but must choose  $x_t$  one period prior to observing  $r_t$  and  $p_t$ . Firms maximize  $E(\pi_t/p_t) = ax_t + dx_t^2 - E(r_t/p_t)x_t$ , and  $E(r_t/p_t) = \alpha + \beta(r_{t-1}/p_{t-1})$ ,  $0 < \beta < 1$ ,  $\text{var}(r_t/p_t) = \sigma^2$ . The sequence  $\{p_t\}$  is independently and identically distributed with mean 1 and is independent of the sequence  $\{r_t/p_t\}$ . Without loss of generality assume that the unconditional expectation of  $r_t/p_t$  is 1, whence  $\alpha + \beta = 1$ . The factor-demand function is  $x_t = (\alpha - a)/2d + (\beta/2d)(r_{t-1}/p_{t-1})$  and the supply function is  $y_t = (\alpha^2 - a^2)/4d + (\alpha\beta/2d)(r_{t-1}/p_{t-1}) + (\beta^2/4d)(r_{t-1}/p_{t-1})^2$ .

Suppose an econometrician estimates the production, factor-demand, and/or supply equations, but assumes that firms make decisions as if previous-period prices would persist in the next period, rather than using previous-period prices to forecast the prices of the next period. If the factor-demand function is assumed to be representative, the econometrician will construct the supply function  $\hat{y}_t = -(a - \alpha)^2/4d\beta + (\beta/4d)(r_{t-1}^2/p_{t-1}^2)$ ; from the production function, he will construct  $\hat{y}_t = -a^2/4d + (1/4d)(r_{t-1}^2/p_{t-1}^2)$ .

Now consider the imposition of an ad valorem output subsidy. Using the methods of the previous section, the average effect of the subsidy on output will actually be  $[(s^{-2} - 1)/4d](1 + \beta^2\sigma^2)$ . But under the assumption of a representative agent the anticipated effect will be  $[(s^{-2} - 1)/4d](2\alpha\beta + \beta^2 + \beta^2\sigma^2)$  using the supply function;  $[(s^{-2} - 1)/4d]\beta(1 + \sigma^2)$  using the factor-demand function; and  $[(s^{-2} - 1)/4d](1 + \sigma^2)$  using the production function. The anticipated effect using the production function overstates the actual effect, on average, while that using the supply function understates it; and except for unreasonably large values of  $\text{var}(r_{t-1}/p_{t-1})$  the anticipated effect using the factor-demand function typically understates the true effect. With an ad rem sub-

sidy, the average effect of the subsidy on output will be  $[(g^2(u) - 1)/4d](1 + \beta^2\sigma^2)$ , with  $g(u) = E[p_t/(p_t + u)]$ . The anticipated effect will be  $(\alpha\beta/2d)[g(u) - 1] + [\beta^2(1 + \sigma^2)/4d][h(u) - 1]$  using the supply function;  $[\beta(1 + \sigma^2)/4d][h(u) - 1]$  using the factor-demand function;  $[(1 + \sigma^2)/4d][h(u) - 1]$  using the production function; with  $h(u) = E[p_t^2/(p_t + u)^2]$ .

## III. Some Numerical Examples

The importance of the representative agent and aggregation problems taken up in this paper can perhaps best be appreciated by constructing some numerical examples. Table 1 provides the ratios of anticipated to actual effects of subsidies when prices are distributed uniformly on  $[1 - v, 1 + v]$ . As one would expect the results are quite sensitive to the size of  $v$ . The value  $v = .01$  is surely a lower bound; .10 may be reasonable for local areas (George Stigler, 1961) while price dispersion over larger areas may be greater. When  $v = .2$ , the highest price is half again as high as the lowest, and when  $v = .5$ , this ratio rises to three. The latter ratio is implausibly large for actual selling prices, but in other, common circumstances may be realistic—for example, when some firms sell retail and others wholesale, or when the effective selling price is anticipated based on information that is not the same for all firms.

The supply function in the ad valorem case is perfectly adequate because an ad valorem subsidy does nothing to distort the distribution of relative selling prices across firms, in our model. When an ad rem subsidy is introduced the distribution of prices is shifted upward, but the aggregate model imputes to the subsidy an increase in the dispersion of prices across firms which in fact does not occur. Since supply is a convex function of selling price, this leads to an underestimate of the effects of the subsidy. As  $u$  increases, this convexity becomes less important, and hence ratios always vary directly with  $u$ . The production function representative agent model performs the worst here because the aggregate production function reflects a compounding of the relationship of output to factor prices, and of factor

TABLE 1—RATIOS OF ANTICIPATED TO ACTUAL EFFECTS, AGGREGATION OVER PRICES<sup>a</sup>

$v$	Supply Function	Factor Demand	Production Function
A. Ad Valorem Subsidy			
.01	1.0000	.9999	.9999
.10	1.0000	.9933	.9867
.20	1.0000	.9731	.9470
.50	1.0000	.8240	.6789
B. Ad Rem Subsidy ( $u/p_t = .20$ )			
.01	.9999	.9999	.9998
.10	.9931	.9864	.9798
.20	.9722	.9461	.9207
.50	.8264	.6809	.5610

<sup>a</sup>Distribution of prices across firms is uniform on  $[1-v, 1+v]$ .

prices to inputs; at each step, the convexity of the relevant function introduces downward bias due to aggregation.

Providing representative examples in the price expectations model is simplified by the fact that results depend only on  $\beta$ , given reasonable values of the other parameters. What is a reasonable value of  $\beta$  depends on the inherent volatility in  $r_t/p_t$ , and the length of the decision period (for a month, .99 might be reasonable; for a year, .75). The outstanding feature of Table 2 is that the production function representative agent model provides an accurate indication of the effects of a subsidy on output, in situations in which the other two models do not. This result emerges because the aggregate production function is not influenced by the expectations process, whereas the other two aggregate models are. The derived supply function is correct for deterministic changes in prices, and the imposition of a subsidy is very nearly that: the only source of discrepancy in this model stems from the use of  $r_{t-1}/p_{t-1}$  rather than  $\beta r_{t-1}/p_{t-1}$  in evaluating the effects of the subsidy on average, and with standard deviations in prices modest relative to their means, that is unimportant.

#### IV. Conclusion

My example has illustrated several facts. Perhaps the most important is the similarity between the dependence of expectations for-

TABLE 2—RATIOS OF EXPECTED ANTICIPATED TO EXPECTED ACTUAL EFFECTS, PRICE EXPECTATIONS<sup>a</sup>

$\beta$	Supply Function	Factor Demand	Production Function
A. Ad Valorem Subsidy or Ad Rem (small $u$ )			
.99	.9999	.9908	1.0008
.95	.9976	.9536	1.0038
.90	.9903	.9066	1.0074
.75	.9389	.7628	1.0171
.50	.7525	.5149	1.0297
B. Ad Rem Subsidy ( $u = .50$ )			
.99	.9904	.9888	.9988
.95	.9593	.9517	1.0018
.90	.9192	.9049	1.0054
.75	.7915	.7614	1.0152
.50	.6270	.5138	1.0277

<sup>a</sup>In all instances  $\sigma = .20$ , and in figuring  $g(u)$  and  $h(u)$  for the right panel the distribution of  $p_t$  is taken to be uniform on  $(.75, 1.25)$ .

mation on policy regime and the dependence of aggregators on policy regime. In the aggregation example with disparate prices across firms, ratios of output prices between specific firms never changed. An ad valorem subsidy preserved that relationship, and hence the aggregator, and so could be evaluated properly using the aggregate supply function from the period with no subsidy. An ad rem subsidy changed the cross-firm output price relationships and the aggregators, and its effect on output was evaluated incorrectly with all representative agent models.

The example shows that it can make a difference *which* representative agent is used, even when aggregation is exact. Aggregators are *not* the same for production, factor demand and supply; the familiar cross-equation restrictions of neoclassical production theory fail and will lead to computational dilemmas. The same problem arises when expectations are ignored. It would be very useful to state with some generality which representative agent is likely to be more reliable in these and other circumstances, but that is likely a difficult undertaking and clearly beyond the purview of this paper.

Finally, these numerical examples show that at least one situation can be constructed in which the perils of ignoring aggregation

are of the same order of magnitude as those of ignoring expectations. Since the example is constructed, any semblance between this finding and difficulties with real data may be purely coincidental. But it does suggest that in the attempt to build econometric models with primitive parameters of tastes and technology, it may be useful to treat aggregation more explicitly than is usually done in these efforts.

#### REFERENCES

- Biddle, J., "The Intertemporal Substitution Elasticity of Labor Supply: Another Look," manuscript, Department of Economics, Duke University, 1984.
- Eichenbaum, M. S., Hansen, L. P. and Singleton, K. J., "Modeling the Term Structure of Interest Rates Under Non-Separable Utility and Durability of Goods," Working Paper, Carnegie-Mellon/GSIA, 1984.
- Lucas, R. E., Jr., "Econometric Policy Evaluation: A Critique," in K. Brunner and A. H. Meltzer, eds., *The Phillips Curve and Labor Markets*, Carnegie-Rochester Conferences on Public Policy, Vol. 1, *Journal of Monetary Economics*, Suppl., 1976, 19-46.
- MaCurdy, T. E., "A Simple Scheme for Estimating an Intertemporal Model of Labor Supply and Consumption in the Presence of Taxes and Uncertainty," *International Economic Review*, June 1983, 24, 265-89.
- Sargent, T. J., "Estimation of Dynamic Labor Demand Schedules under Rational Expectations," *Journal of Political Economy*, December 1978, 86, 1009-44.
- \_\_\_\_\_ and Hansen, L. P., "Formulating and Estimating Dynamic Linear Rational Expectations Models," reprinted in R. E. Lucas, Jr. and T. J. Sargent, eds., *Rational Expectations and Econometric Practice*, Minneapolis: University of Minnesota Press, 1980.
- Stigler, G. J., "The Economics of Information," *Journal of Political Economy*, June 1961, 69, 213-26.

## MODELING INTERCOUNTRY LINKAGES<sup>†</sup>

### An Integrated Accounting Matrix for Canada and the United States

By JACOB COHEN AND STEVEN HUSTED\*

In the past two decades, considerable progress has been made in studying the economic relationships between countries through the linkage of large-scale national econometric models. Examples of these projects include Project Link, the RDX2-MPS experiments of the Bank of Canada, and the multicountry model of the Federal Reserve Board. While these models tell us much, they typically suffer from several important problems. First, the linkages are often incomplete. In some cases the separate country models may only be linked via the trade accounts. Or, if capital and factor flows are considered, they are modeled in only a highly aggregated fashion. Second, as Ray Fair (1979) has noted, no model does an adequate job of linking the underlying sectoral flows of funds accounts with the national income accounts. Therefore, in the interest of modeling aggregate relationships, underlying balance sheet constraints may be violated or ignored. This could have, Fair argues, important consequences for empirical results.<sup>1</sup>

In this paper we suggest our own strategy for modeling the economic linkages between any two countries. Our strategy focuses on the underlying flows between these two countries and the sectoral contributions to these flows. That is, we propose to merge the flow of funds accounts via their bilateral balance of payments. We envision an integrated flow of funds accounting framework with the two countries sharing a common balance of payments. This modeling strategy has several advantages. First, our suggested framework could be used, following the strategy of Fair, to supply the financial underpinning for future economic modeling of international linkages of prices and interest rates. The balance sheet constraints inherent in the framework will impart additional information in any statistical estimation of such a model. Second, this strategy should yield a better understanding of a country's balance of payments since it necessarily links domestic decision making with its international outcome. Third, our framework should be useful to policymakers since it would allow them to model the underlying financial implications of proposed policy changes. Finally, the implementation of our proposal could lead to improved accuracy in balance of payments statements.

<sup>†</sup>*Discussant:* John A. Sawyer, University of Toronto.

\*Department of Economics, University of Pittsburgh, Pittsburgh, PA 15260. Too numerous to mention are the individuals who have offered suggestions and statistical assistance. Two who must be singled out, however, are Gerry Gravel and John Motala of Statistics Canada. We are also indebted to Douglas Case for research assistance. Financial support was provided by the University Center for International Studies, University of Pittsburgh and the Canadian Embassy, Washington, D.C.

<sup>1</sup>Fair demonstrates this point via a hypothetical example. In this example, he duplicates his own econometric model of the United States which includes sectoral financial balance sheet constraints for four economic sectors. The sectors Fair considers are households, firms, banking, and government. Then, he links the two models

so that his "world" is composed of two identical countries wherein all flows of funds between the two countries are accounted for and the exchange rate is endogenously determined. Using this model, Fair goes on to simulate the effects of a variety of economic shocks under varying assumptions about capital mobility and exchange rate regime.

### I. Conceptual Framework

Both the United States and Canada maintain and report detailed flow of funds (*FOF*) accounts. The Federal Reserve is responsible for constructing the U.S. accounts. In Canada, the accounts are prepared by Statistics Canada. The lack of Bank of Canada involvement creates data problems.

The *FOF* accounts report financial and nonfinancial activities for various sectors and subsectors of the economy (for example, households, business, government, commercial banks, nonbank financial institutions, monetary authority, rest of the world). The classification is institutional rather than functional. That is, the sectors are decision-making sectors. In the business sector, where this matters most, corporations are the organizational unit rather than operating plants. If establishment-institutional matrices could be constructed, production decisions could be linked with the financial decisions of the corporation. Moreover, this would permit disaggregation of the business sector so that interindustry flows could be recorded. To date, such linkages have not been achieved for either the United States or Canada.

The summary transactions matrix prepared in both countries shows sectors down the columns and transactions across the rows. Nonfinancial transactions are reported first and consist of saving entries (credits) and capital outlays (debits) derived from the national income and product accounts. Next follow the financial transactions rows which include detail on financial uses (debits) and sources (credits) of funds or in an alternative (Canadian) terminology, detail on increases in financial assets and liabilities. There are important differences in transaction and sector coverage in the two sets of accounts. Consumer durables are included in U.S. capital outlays, but not in Canadian. Government capital acquisition and net purchases of existing assets are included in Canadian nonfinancial capital acquisitions but not in the U.S. acquisitions. In the financial rows, claims on associated enterprises are part of the Canadian accounts, but not the U.S. accounts. Also, the Canadian accounts break out public financial institutions, social secur-

ity funds, nongovernment enterprises from the government sectors. Persons and unincorporated business are combined in the Canadian data, but are separated in the U.S. accounts.

Across all sectors, saving and capital outlays are equal. For individual sectors, the identity is between gross investment (capital outlays and net financial investment) and gross saving. Net financial investment (financial uses less financial sources) over all sectors is zero.

In addition to these flow of funds constraints, an integrated two-country matrix introduces a set of *FOF* identities based on the bilateral balance of payments. Once transactions are stated in a common currency, the debit (credit) items for the U.S. subsector of the Canadian rest of the world sector will have a counterpart in the credit (debit) items of the Canadian subsector of the U.S. rest of the world sector. In schematic form, an integrated matrix will enter the second country sectors in reverse order so that one rest of the world sector and its mirror image lie alongside each other.

The *FOF* accounts are subject to various kinds of statistical errors and inconsistencies: in valuation bases, in timing, in classification of transactions and omissions of pertinent transactions or transactors from the accounts. For the *FOF* constraints to hold, balancing statistical discrepancy rows and columns must be introduced. The sum of sector discrepancies (called the residual error of estimate in the Canadian accounts) equals the sum of transactions discrepancies. A large statistical discrepancy for the household (persons and unincorporated business) and rest of the world sectors is a feature of both sets of accounts. A strong correlation has been found in the U.S. accounts between these discrepancies. Errors in the capital flows estimates have their counterpart in estimates of household acquisition of financial assets (Frank de Leeuw, 1984). A reduction in statistical discrepancy for the rest of the world sector by reconciliation of the Canadian and U.S. balance of payments may improve household estimates and contribute to a reduction in overall discrepancy for both countries.

Despite apparent differences in the calculation of statistical discrepancy, Canadian and U.S. accountants looking at the same data will arrive at identical discrepancy measures. This is because statistical discrepancy is uniformly defined in terms of the inequality of gross saving (*GS*) and gross investment (*GI*) (over all and for each sector). Gross investment is the sum of capital outlays (*CO*) and net financial investment (*NFI*). Hence, statistical discrepancy (*SD*) is calculated according to the following formula:

$$SD = GS - GI = GS - CO - NFI.$$

Over all sectors, the U.S. statistical discrepancy reflects both the discrepancy in the national income accounts and the discrepancy in net financial investment. In the Canadian accounts, in contrast, financial debits and credits for given transactions are forced into balance so that net financial investment is zero. This process requires that the Canadian statistician reestimate gross saving or capital outlays. The end result of this adjustment is that the original discrepancy between *GS* and *GI* will be left unaffected.

## II. Estimation Problems

Our goal is to merge the two *FOF* accounts into a single matrix. In order to achieve this goal, a number of estimation problems must be overcome. First, it is necessary to have a uniform transaction classification for the two countries. The financial transactions lines necessarily distinguish between Canadian, U.S., and rest of the world financial instruments. A Canadian instrument can be bought or sold in all three markets and the same for U.S. and third-country instruments. In some cases, transactions do not match up across countries (for example, claims on associated enterprise) necessitating consolidation of categories.

Unlike the necessary uniformity of transaction coverage, institutional sectoring can reflect national differences. The only required matching is in the rest of the world sector, which will now be split into U.S. and

third-country sectors for Canada, and Canada and third-country sectors for the United States.

Breaking the foreign sectors of the two *FOF* accounts into the "second" country (i.e., the U.S. or Canada) and the "rest" of the rest of the world poses one of the major challenges to our modeling effort. This is because this breakdown is not made in the published *FOF* accounts of either country. Consequently, we must look to other sources such as each countries' bilateral balance of payments for the necessary data. But the detailed categories in the balance of payments lack a one-to-one correspondence with the *FOF* transactions. Assigning U.S. figures in the Canadian balance of payments to the *FOF* transactions categories produces residual rest of the world values in the *FOF* that have no resemblance to the counterpart third-country flows in the Canadian balance of payments. Statistics Canada kindly provided a reconciliation of *FOF* and the balance of payments with the U.S. broken out for 1981 but without some methodology a continuous time series for the U.S. subsector could not be constructed. We skirted this problem by forecasting only three main aggregates for the rest of the world subsectors—saving, capital outlays (transactions in existing intangible assets), and net financial investment. This seemed to provide reasonable consistency between the *FOF* accounts and the balance of payments data on which we had to rely.

A second major challenge in the use of balance of payments data is reconciling the statements of the two countries. Both the United States and Canada have long published detailed bilateral balance of payments statements. For a variety of reasons, the reported data on major line items seldom match and, in fact, the mismatches are often very large. Rising concerns in the 1960's over this problem led to the creation of the Technical Working Group on Canada-U.S. Balance of Payments in 1963. Annual reconciliation of the current account during 1968-71 was based on the work of this committee and since 1971 on that of the United States-Canada Trade Statistics Committee. Their detailed reconciliation of trade flows in 1974

reveals three main kinds of discrepancy—geographical classification, transactions classification, and a catchall category which includes nonreceipt of documents, timing, and valuation differences and errors (Statistics Canada, 1981, p. 301). The adjustments were large enough in 1969 so that the U.S. perception of its current account position with Canada shifted from a deficit to surplus.

Reconciliation figures have yet to be produced for the capital accounts. Differences in the amount of detail are one difficulty. Problems are in part conceptual such as with the treatment of U.S. branches of Canadian insurance companies. Both the United States and Canada treat these branches as nonresidents. A study for 1972 pinpoints differing estimates of short-term capital flows as the main source of the bilateral discrepancy (Statistics Canada, p. 308). The “third-country” problem also contributes to errors in the bilateral data. Such errors arise from settlement via third countries (“multilateral settlement”). For example, a Canadian importer pays a European exporter with U.S. dollar balances. As a result the bilateral balance of payments for all three countries will be thrown out of balance.

The difficulty in reconciling the bilateral accounts is perhaps best illustrated by comparing the signs on the errors and omissions entries in the two bilateral accounts. An “outward” discrepancy (net debits) for one country is not matched by an “inward” discrepancy (net credits) for the other country. Over the period 1946–74, the annual signs were the same in 17 of 29 cases (Statistics Canada, pp. 166–67). Reconciliation of the capital accounts is a separate research exercise. For the present we are relying on the more detailed Canadian data in estimating the integrated matrix.

### III. Forecasting the Matrix

Once we had a feel for the contours of the matrix, we set about to forecast the cells for the year 1985. A possible first step in constructing such a forecast would be to examine the existing large scale models of the United States and Canada for clues as to

how to model statistically nonfinancial and financial flows. If one were to use these equations and then rely on judgment and the adding-up constraints inherent in the model, then the only additional information required would be a set of predicted values of the exogenous variables of the system.

An alternative procedure and the one we employed in our work is to forecast the matrix using predictions generated from simple time series models of the data. At first, we forecast both detailed component entries of the *FOF* accounts and the aggregates of these components. This led to the problem that the sum of the forecasts of the component parts did not equal the forecast of the aggregate. Faced with the inevitability of a proliferation of discrepancy balancing items, we used only the forecasts of the main aggregates for the integrated matrix. These aggregates were gross saving, nonfinancial capital acquisition (outlays), net financial investment, net increase in financial assets, net increase in liabilities, and statistical discrepancy. (As previously stated, only three aggregates were forecast for the rest of the world sector.) Forecasting net financial investment in addition to its components provided a way of residually estimating either financial uses or sources when the predicted value was shaky. For each sectoral aggregate, simple distributed lag equations containing alternately one and two lags were fitted to the historical data. Because the aggregated data often evolve slowly over time, this exercise commonly yielded prediction equations with  $R^2$  exceeding .90. Credence to this approach is given by the often recorded success of time-series methods in economic studies (albeit more sophisticated than our own) over other more complicated forecasting techniques. A major problem with this technique, however, is that it is incapable of predicting major turning points in the data. Because of this, once such a turning point has been missed, this error is factored into subsequent forecasts.

A point of interest was the general deterioration in our results when we applied our forecasting methodology to changes in individual assets and liabilities. An implica-

tion of this is that the larger magnitudes such as capital outlays are flow constrained (the lagged values are proxies for income or external finance), while detailed portfolio decisions are price (interest) constrained.

Once the forecasts had been generated, they were entered into the matrix. Here judgment was required since the adding-up constraints were necessarily violated by the independent forecasts. At least one of the forecast values would have to be adjusted for each sector and each row. An iterative procedure in adjusting the rows and columns of the matrix was followed. The direction of adjustment was in line with the historical trends in the series. This method is similar to the method of successive approximations employed by the Federal Reserve in its flow of funds projections. Our application of this technique simultaneously to two countries is a point of departure. Another point of departure from the Fed's procedures was our recognition of the inevitability of statistical discrepancy and treating it as another variable.

We found in constructing our forecasts that frequently the discrepancy forecast was more reliable than the individual financial and nonfinancial transaction forecasts. This was especially true of the discrepancies in the U.S. *FOF* accounts where we obtained  $R^2$  values exceeding .6 in seven of fourteen sectors. When we estimated equations for rest of the world discrepancies for both countries we found that interest rate differentials, trade flows and the level of direct foreign investment contributed significantly to explaining these variables. (For more detail on this point see our 1984 paper). Hence, instead of treating the discrepancy as zero *ex ante*, other flows could be predicted residually once the discrepancy had been forecast. In those cases where we were not confident of our discrepancy estimate, it was generated residually.

While our forecasts cannot be reported fully here, an inkling of what they are like is suggested by our forecasts of the disaggregated rest of the world subsectors for Canada and the United States for 1985. (In constructing our forecasts we used annual data

for the period 1950–83 for the United States and, 1960–83 for Canada.)

Expressed in millions of U.S. dollars (on the basis of an exchange rate of .79), we forecast the following for the U.S. subsector of the Canadian matrix (numbers in parentheses are 1983 values): gross saving (credit entry), -616 (1578); nonfinancial capital acquisition (debit), -46 (-50); net financial investment (debit), 408 (1240); discrepancy (debit) -978 (388).<sup>2</sup> (The last value was forecast rather than residually determined.) Taken together, these values imply that Canada will have a current account surplus with the United States but, because of the size of the forecast discrepancy, the United States will have a positive net financial investment in Canada. For the foreign rest of the world subsector for Canada we forecast: gross saving, 771 (-2214); nonfinancial capital acquisition 241 (676); net financial investment 4592 (2748); discrepancy -4062 (-5638). For the foreign rest of the world subsector for the United States we forecast: gross saving 68935 (33875); nonfinancial capital acquisition -46 (-50); net financial investment 68787 (31543); and discrepancy 204 (2382). Taken as a whole, our forecast of the 1985 U.S. balance of payments suggests a current account deficit of about \$70 billion, approximately twice the 1983 level.<sup>3</sup>

#### IV. Future Work

The work to date has been an exercise in filling the cells of an integrated two-country *FOF* matrix. We do not have a great deal of

<sup>2</sup>Recall that these numbers will appear with the opposite sign in the Canadian subsector of the U.S. portion of the matrix. Nonfinancial capital acquisition in the *FOF* rest of the world sector for Canada is limited to "inheritances and immigrants' funds" as reported in the Canadian balance of payments data. Receipts and payments were netted against each other vis-à-vis the United States with the residual assigned to the "foreign" sector. The U.S. *FOF* matrix does not record such transactions.

<sup>3</sup>The U.S. rest of the world forecasts were generated under the assumption of a 1984 current account deficit of \$100 billion rather than using the forecast deficit for 1984, based on the autoregressive equations.



confidence that estimation by autoregressive models and subsequent iteration will provide good forecasts of either 1984 or 1985. But, it is a beginning for more sophisticated work. A possible next step would be the fitting of simultaneous econometric equations to provide a genuine flow of funds behavioral model. The end result would be a structure similar to Fair's, but with a more-detailed financial underpinning. We caution however that equations based on flow of funds accounts have special problems. Outside the estimation period the tracking performance is often unsatisfactory. The fragility of financial equations has many explanations—financial innovation, the speed with which financial markets can adjust to changes in rates, and most importantly the inadequate modeling of market expectations. Therefore, any forecasting effort should follow a modest approach which combines the discipline of the flow of funds with "judgment" and selective econometric estimation.

Until reconciliation of the bilateral capital accounts, model building of the type we have proposed will probably have to rely upon one set of balance of payments data and impose it on the second country. Reconciliation has to be a government affair although the outside economist can be of use in outlining the conceptual problems that would be

faced and suggesting testable hypotheses. If the Canadian and U.S. governments can be encouraged to tackle the discrepancy problem, a positive result would be the publication of a unified set of data on bilateral transactions. Such cooperation if copied by others should throw light on errors and omissions in the balance of payments on a global basis—a matter of considerable interest and concern.

## REFERENCES

- Cohen, Jacob and Steven Husted, "An Integrated Accounting Framework for Canada and the United States," paper presented at the Canadian Economic Association Meetings, Guelph, Ontario, May 1984.
- de Leeuw, Frank, "Measuring Private Saving," mimeo., U.S. Department of Commerce, Washington, 1984.
- Fair, Ray, "On Modelling the Economic Linkages among Countries," in R. Dornbusch and J. Frenkel, eds., *International Economic Policy*, Baltimore: Johns Hopkins University Press, 1979, 209–38.
- Statistics Canada, *The Canadian Balance of International Payments and International Investment Position: A Description of Sources and Methods*, Technical Report, Statistics Canada, 1981.

# Modeling U.S.–Mexico Economic Linkages

By CLARK W. REYNOLDS AND ROBERT K. MCCLEERY\*

The United States and Mexico share a unique history and common border. Despite efforts by both countries to regulate the degree of exchange between them, relentless economic forces give rise to what has been called increasing “silent integration” in recent years. It has become apparent in this and similar cases that trade barriers tend to exacerbate migration, that capital flows in one direction can moderate labor flows in the other, and that limitations on direct investment may lead to even more vulnerable patterns of international indebtedness. These pressures for economic integration operate regardless of policymakers’ preferences, posing a challenge in two separate but related areas of economic analysis. First, by having to look at the mutual interaction of trade and factor flows (including labor, capital, and technology transfers), the analysis of exchange is pushed well beyond its conventional limits to accommodate wage, income, and growth effects in sending and receiving countries. In such a context, normal trade theory appears even more partial and its limitations all the more restrictive. Dynamic links between trade and factor movements and growth and structural change in the respective economies cannot be ignored in cases where initial policy decisions on the part of one country may have a major impact on both, with feedback effects that stretch the limits of intuition.

The second area of analysis challenged by these issues is that of political economy. Here one wishes to assess the costs and benefits to

relevant social groups in both countries of alternative trade, migration, investment, and technology patterns arising from a range of policy options. Such an analysis is complicated not only by the challenges it imposes to the economics of exchange, as mentioned above, but by the fact that in this case the political and social systems of both countries are so different. Partly because of these differences, the flows between them remain highly asymmetrical in terms of benefits, costs, and power relations. Tradeoffs between exchange in goods and services and movements of labor, capital, and technology are further complicated when they contribute to the dynamic pattern of comparative advantage in both countries, as is the case in the United States and Mexico. Moreover the problem increases when both partners face the need for sustained fiscal adjustment, when exchange rates in both countries have been distorted by past policy imperfections, and when the rest of the international economic system is involved in an uncertain path of evolution in production, employment, and technology.

## I. Managing Interdependence

The tension between the benefits from a fuller degree of exchange and the costs perceived by groups on either side of the border leads to a classic political economic problem of managing interdependence. The analysis of interrelated trade, migration, and investment flows in this case cannot only benefit from different but comparable examples elsewhere, such as the integration process among European countries at different levels of development, but can also shed light on ways to model the economic dimensions of policy space open to decision makers. By showing the tradeoffs more clearly, and by assessing their economic impact on different interest groups, the analysis sharpens the range of policy options. There is no attempt in this paper to go beyond a relatively simple

\*Professor of Economics, Food Research Institute, Stanford University, Stanford, CA 94305, and graduate student in Economics, Stanford University. We acknowledge substantial computer assistance from Robert O. Wood, graduate student in Economics at Stanford University. Support for this research was provided by the U.S.–Mexico Project, Stanford University, under grants from the William and Flora Hewlett Foundation, the Andrew W. Mellon Foundation, and the Rockefeller Foundation.

dynamic model of exchange between two interdependent economic systems, based on stylized facts drawn from the recent experience of the United States and Mexico. Our simulation model (see our forthcoming working paper) assumes some degree of competition in a capitalist-mixed enterprise system subject to reasonable patterns of accumulation, growth, and employment under five different policy scenarios: (i) the status quo (likely flows based on current migration and investment policies); (ii) much freer movement of labor and capital; (iii) increased labor market controls (Simpson-Mazzoli-type migration policy); and managed interdependence affecting (iv) capital and (v) labor flows. Scenario *v* involves the imposition of a "migration tax" similar to that recommended by the Kindleberger Commission in its OECD report (Charles Kindleberger et al., 1978).

## II. A Migration Tax

There are important differences between the migration tax proposal developed by the Commission and that simulated here, reflecting major dissimilarities between the predominantly legal migration flows in Europe and the undocumented, so-called "black market" labor flows in the North American case. However, the basic purpose of both proposals is similar: to reduce the level of migration, offset some of the cost that migration imposes on social groups in the receiving country, and provide employment-generating transfers to the sending country that will reduce future migration pressures.

The illegal nature of most North American migration leads to a differential in wages of unskilled labor between the two countries that is significantly in excess of the amount required to cover costs of transportation, job search, inherent language and skill differences, and nonmonetary costs of relocation in the United States. We shall call this the "undocumented differential" in wages. Part of this differential reflects economic rents accruing to legal migrants and U.S. workers owing to labor market protection. Part reflects the cost of "coyotes" who smuggle workers across the border. The remainder is the return to those who exercise monopsony

power over undocumented workers. Legalization of migration would eliminate the cost of coyotes, erase exploitation of migrants due to their undocumented status, and reduce the risks associated with migration. Thus the annual level of a "neutral" tax would be equal to those benefits of legalization, which we have roughly estimated as \$2,000. Any specific proposal would, of course, be subject to negotiation based on more refined estimates of the undocumented differential. By the same token, the distribution of tax proceeds between the two countries should also involve binational negotiation, while the split within each country would be a matter of domestic concern.

In the European case, a tax was proposed that would directly reduce the incentive to migrate. A portion of the revenues would be used to encourage those already in the high-wage country to return. In the case of the United States and Mexico, however, the migration tax could be, in effect, a "windfall rents tax" on the benefits of legalization of migration flows. Because the tax would be levied only on the newly legalized migrants, it would leave the initial incentive to migrate and the stock of migrants unchanged. Over time, however, sharing of tax revenues with Mexico can be shown to significantly reduce the incentives of workers to migrate from the supply side.

## III. The Analytical Framework

The approach taken is to apply a simple model of dynamic wave convergence in each one-year period, subject to assumed cost, skill, and preference differentials between the two countries as described above. The conditions that tend toward convergence include relative technological progress and relative increases in the capital-labor ratio (increases in the capital stock and/or decreasing demographic pressures) in the low-wage country, as well as increased exchange between the two countries under reasonably competitive conditions. This can be due to patterns of development in either country or both. In our model, movements of labor, capital, and technology between countries in response to differential rates of return not only decrease these differentials but also affect growth rates

in the respective regions. Increased exchange refers to flows of labor, capital, and technology as well as trade in goods and services. In the formal model, changes in the degree of trade protection are not explicitly introduced, except implicitly through the accommodation of joint production for both markets. Such accommodation will necessarily underlie capital transfers of the magnitude discussed here.

In order to capture some elements of the process of wage convergence (or divergence) we have employed a two-sector model for the United States and Mexico. Each economy has a high-wage, relatively capital-intensive sector and a low-wage sector with relatively low capital intensity. One mechanism of wage convergence is through migration from Mexico's low-wage sector to that of the United States. Another mechanism is through capital flows from the United States to Mexico in response to differential rates of return net of risk, subject to foreign investment barriers. The technology transfer process will be explicitly modeled in our forthcoming version. At present it is handled through exogenous sectoral productivity growth assumptions.

The sectors are defined as follows. The low-wage sector in the United States (*US2*) represents low-skilled labor including personal services, construction workers, and industrial and agricultural laborers. In the base year 1982 this constituted about 10 percent of the U.S. labor force (10 million workers) and represents those workers with which Mexican migrants are in direct competition. The remaining 90 percent is defined as the high-wage sector in the United States (*US1*). The high-wage sector accounted for 92 percent of the capital stock and 95 percent of value-added of the United States. The low-wage sector in Mexico (*MEX2*) represents all primary and tertiary employment (including agriculture and services) constituting 75 percent of the occupied work force (17 million workers) in the base year 1982. The remaining 25 percent is comprised of those in secondary sector (*MEX1*) employment (primarily manufacturing but including oil) which accounts for one-third of both the capital stock and value added.

In our analysis, standard Cobb-Douglas production functions are used to model pro-

duction in each sector. A feature of our model is that gross national income is allowed to differ from gross national product by net factor payments abroad. A wage differential between sectors within each country is maintained by introducing proxies for union power, demand constraints, and a limit on the substitutability of labor for capital which we call a minimum incremental capital-labor ratio. Between the low-wage sectors in Mexico and the United States, the following equilibrium condition holds: wages in *MEX2* are equal to wages in *US2* less the disutility of migrating which comprises the following five variables: cost, preference, skill, discrimination, and tax. Cost is the approximate monetary cost of getting to and across the border and finding a place of employment in the United States. Preference reflects the fact that, holding incomes constant, Mexicans prefer life in Mexico to that of the United States. It is a proxy for family ties, climate, relative price and availability of normally consumed goods and services, and national differences in consumption bundles. The skill variable accounts for differences in education, training, and language capabilities. The discrimination variable accounts for the costs of illegality including the exercise of monopsony power by U.S. employers. The tax variable consists of percentages of best-judgment figures for cost, preference, and discrimination totalling \$2,000, or approximately one-fourth of the real wage differential for unskilled labor between the two countries.

The model works recursively. For a given year, capital is allocated between sectors in each country such that its rate of return is equalized across sectors. Then wages are computed and labor migration takes place in response to absolute real wage differentials between the respective low-wage sectors. In an iterative fashion, capital is subsequently reallocated within each country, and the postmigration labor force is also reallocated between sectors to some extent, reflecting constraints representing labor market segmentation and skill differentials. Wages are recalculated and compared at this new partial equilibrium, and this process is repeated until the wage differential (modified by the cost, preference, skill, discrimination, and in

one case, tax variables) is effectively zero. Then other macroeconomic variables are calculated and recorded, exogenous variables are updated to reflect the passage of another year, and the process begins again for each year from 1983 until 2000.

In the model the exogenous variables for which growth rates are specified are the indigenous supply of labor in Mexico and the United States, as well as pure productivity growth in each sector of the two countries. Given these assumptions, the model is used to simulate the paths of the endogenous variables which include both countries' sectoral and total *GDP*, wage levels, returns to capital, national income, net investment, stock of migrants in the United States, and the allocation of capital and labor between sectors for each of the five scenarios. This permits a range of migration and investment policy options to be studied. There are two interesting sources of feedback: the effect of increases in the rate of return to capital in the United States due to migration or capital transfers on investment in the United States; and, in the migration tax case, the effect of migration on tax revenues, capital stocks, wages, and migration in future periods. We have arbitrarily assumed one specific distribution of the migration tax revenue, 50-50 between the United States and Mexico, and 50-50 between labor and capital within each country. This significantly alleviates both the opportunity cost to U.S. workers of a lower increase in real wages relative to autarky and the cost to capital owners of a lower level of capital income relative to the status quo. In Mexico the disbursement of tax revenues aids both growth (increases in the capital stock) and income distribution (wage subsidies for low-skilled labor).

#### IV. Some Initial Findings

(i) *The Status Quo*: The initial results of our model indicate that the total stock of undocumented Mexican migrants would follow a bell-shaped path over time, rising from an estimated 2.5 million in 1982 to a peak of about 5 million in 1995, associated with *GDP* growth rates of 5.5 and 2.8 percent, respectively, in Mexico and the United States over the entire period through 2000. Between the

two years (1982 and 2000) the absolute wage gap for low-skilled labor between the two countries remains unchanged. In 1982, wage levels averaged \$10,100 and \$1,900, respectively, while at the year 2000, the model estimates them to be \$10,550 and \$2,350, reducing the relative gap slightly from over 5 to 1 to less than 4.5 to 1. All figures given are 1982 U.S. dollars. Real *GDP* rises from \$3,000 billion and \$150 billion in 1982 in the United States and Mexico, respectively, to \$4,940 billion and \$390 billion in the year 2000. (These figures take into consideration the slump in 1982-84, while assuming a fairly steady upward trend thereafter in both countries.) Real wages in the high-skilled sector of the United States and Mexico in 1982 averaged \$24,350 and \$3,680, respectively, rising to \$31,140 and \$6,130 at the end of the period. The relative gap in high-skilled wages declines from 6.6/1 to 5.1/1, a much larger decline than in the case of low-skilled labor. If these trends in migration and wage levels are unacceptable, alternatives to the status quo should be considered. These figures provide the base line against which we present the results of the other scenarios.

(ii) *Free Movement of Labor after 1985*: This scenario is not presented as a realistic policy option, but as a boundary case which sets one limit to the economic dimensions of policy space. Assuming that Mexican workers exhibit a constant preference for living in Mexico and that relocation costs do not rise, the elimination of legal barriers to migration would in this extreme case also produce a bell-shaped curve in the stock of migrants which peaks only after the year 2000. In the first year of legalization, nearly 8 million Mexicans migrate, and the stock reaches the unrealistic level of 16.5 million. Even then there is not full convergence of real wages of unskilled labor which become \$10,460 in the United States and \$3,260 in Mexico. (Of course, long before that point social and political reactions in both countries would have put a lid on the flow that could be far more destabilizing in the long run than the current "black market" in labor flows.)

(iii) *Increased Labor Market Controls*: The Simpson-Mazzoli alternative is simulated with some difficulty, due to ambiguities in just how employer sanctions and increased

border patrol funding translate into our cost variable. First and most importantly, we find that even if Simpson-Mazzoli were to eliminate all undocumented migrants in the first year, the flow may well resume leading to a bell-shaped stock of migrants that would peak in the late 1990's at a level over 1 million more than those legalized at the outset. The process of relative wage convergence would be set back. Simpson-Mazzoli is, of course, only a step toward the other limit of economic policy space, namely autarky (prohibitive barriers to migration). Full autarky would maximize the protection rents to U.S. labor and minimize the degree of convergence through migration. The U.S. wages for unskilled labor would rise by \$40 over the status quo projection while those in Mexico would reach only \$2,200 in the year 2000 compared to the status quo projection of \$2,350, an 8 percent loss. Autarky would also build up pressures for capital to move offshore (though not necessarily to Mexico) with adverse consequences for U.S. wage levels. In the Mexican case, the growth of real wages is sharply reduced. The resources devoted to enforcement of such a restrictive regime would be immense, perhaps exceeding the benefits to workers in the United States, and capital owners in the United States would lose in any event.

(iv) *Managed Interdependence of Capital Flows*: This proposal calls for a significant increase in the amount of co-investment for joint production between the United States and Mexico to eventually serve both markets, with a 50-50 sharing of the proceeds. It is assumed that such investment would neither replace nor stimulate Mexican domestic capital formation, though the latter is likely to result. Migration policy in this scenario is presumed to remain unchanged. The flow of \$5 billion per year between the two countries beginning in 1986 proves to have a significant impact on migration flows, so that the stock of Mexican migrants in the United States peaks in 1985 at 3.5 million (1.5 million below the status quo peak of 5 million in 1995). By the year 2000 this scenario leads to a total stock of migrants of less than 1 million, implying a return flow over the period from 1986 to 2000. Real wages of unskilled labor increase by \$50 in both coun-

tries relative to the status quo scenario, despite a much larger number of workers remaining in Mexico. The capital flow case leads to skilled labor wages which are \$80 higher in the United States and \$150 higher in Mexico than the status quo scenario. Total gross national income in both countries combined is higher than in the status quo case by \$60 billion 1982 dollars in 2000, with the gains split roughly equally. These findings are indicative of the power of managed interdependence in the investment area.

(v) *Managed Interdependence of Labor Flows through a Migration Tax*: The implementation of a tax of \$2,000 would also result in an earlier and lower peak in migration than the status quo, at a stock of about 3 million migrants in 1993. In comparison with the status quo Mexican workers gain, Mexican capitalists gain (the Mexican capital stock grows more rapidly due to the tax and income feedback effects so that the declining return to capital does not slow growth), U.S. capitalists lose slightly as the number of Mexicans in the United States falls below the status quo level, while U.S. workers gain. Combined gross national income of the two countries increases by \$25 billion in the tax case, with a Mexican share of \$16 billion.

In a final scenario the two types of interdependence management, tax and investment policies, are combined. This leads to a fall in the stock of migrants to nearly zero in the year 2000, an increase in joint income of \$70 billion over the status quo in the final year, with the benefits split evenly between the two countries. Real wages of unskilled labor are about \$60 higher in both countries than in the base case, giving a slight improvement for each over the case of investment policy alone.

## V. Conclusions and Implications for Further Analysis

This study illustrates the positive economic impact which Mexico and the United States could receive from some degree of managed interdependence. A migration tax could make use of the gains from "rationalizing" migration policy in a way that compensates workers in the United States for wage gains foregone, significantly reduces the

flow of migrants, and serves to improve income levels and distribution in Mexico. There is an identifiable tradeoff between U.S. investment in Mexico and migration to this country. In our model an annual flow of \$5 billion would reduce the stock of migrants in the United States by 200,000 per year on a cumulative basis. An advantage of the combined tax and investment proposal is that with only 20 percent of the migration that would obtain in the absence of all immigration barriers, 70 percent of the benefits would be achieved in terms of joint income with a more equal binational distribution of the benefits.

In all scenarios, Mexican migration peaks within a generation and in most cases before 1995. These findings strongly support the contention by Mexican sources (F. J. Alejo, 1983) that the migration problem for them is one that is greatest before 1995, as their own economy becomes increasingly able to absorb its own labor, while the supply of new workers decelerates reflecting sharp fertility declines which are already in evidence. The United States, on the other hand, will exhibit increasing demographic complementarities with Mexico as its population ages, requiring more services, even as the proportion of those willing to enter the lower echelons of the labor market declines.

Despite fundamental differences between the U.S.-Mexico case and other experiences,

there are important similarities reflecting what might be called the underlying international development problem. This problem affects all countries, rich and poor. If global political economic stability is to be maintained, the problem must be addressed by policies that serve to manage interdependence in ways which minimize costs and uncertainties and spread benefits as widely as possible within and between countries. For interdependence to be managed at all well, it is useful to engage in some modeling of dynamic economic linkages among countries along with the social and political consequences of alternative policies. Hence the U.S.-Mexico case has potentially wide applicability, perhaps as much in raising relevant questions as in answering them.

#### REFERENCES

- Alejo, F. J., "Mexican View of the Political-Economic Implications of U.S.-Mexico Trade and Financial Interdependence," in edited papers on U.S.-Mexico Trade and Financial Interdependence, U.S.-Mexico Project, Stanford University, September 15-17, 1983.
- Kindleberger, et al. C.P., *Migration, Growth, and Development*, Paris: OCED, 1978.
- Reynolds, C. W. and McCleery, R. K., "U.S.-Mexico Project," working paper, Stanford University, forthcoming Spring 1985.

# New Developments in Project LINK

By L. R. KLEIN\*

An attempt to model the international transmission mechanism was initiated by project LINK in the summer of 1968. There have been many interim reports on the progress of that effort, over the years, the last major account being some three years old. (See my paper with Peter Pauly and Pascal Voisin, 1982.) It is accordingly appropriate to bring the story up to date.

In the early years of this international cooperative research project, the Bretton Woods system of fixed parities was still alive, but the shift to floating rates presented an immediate research challenge to provide endogenous generation of exchange rates on the basis of meager experience and short data samples. The change in the terms of trade between oil importers and oil exporters provided another challenge, but fortunately the system was constructed from the start with explicit display for energy trade.

Apart from meeting these challenges and coping with the more ordinary changes that are always occurring in the dynamic economy of the real world, the project researchers have always sought to enhance the system. The international model began as a system that was built around trade flows among the leading OECD countries, some thirteen of them. We saw immediately the special problems of model building for developing countries and for centrally planned economies. It was immediately apparent that we would have to include the whole world in the system because the LINK concept was to preserve world accounting identities: world-exports = world imports. These are plural identities because they must hold by commodity groups and in both real and nominal terms. We resolved the issue by programming the computations in a specific way so that each country's exports = weighted sum

of partners' imports and import prices = weighted sum of partners' export prices.

The global *consistency* conditions were achieved in the LINK system by programming techniques that have been explained in detail in previous project summaries. Over the years, other international model-building projects have been implemented in various government agencies, multinational agencies, research consulting enterprises and some universities. There is now a whole corps of international models; some emphasizing trade, some emphasizing financial flows, and some emphasizing partial interrelationships.

A feature of the LINK system, however, has been its emphasis on the principle that each resident model builder knows his or her own country (area) best. LINK, therefore, remains as a cooperative venture, with participants from model-building centers round the world. Models are maintained for separate use at home but also sent to LINK central under certain specifications for consistency at linkage points, and world-consistent simulations are made for the project as a whole. The cooperating researchers meet, face-to-face, twice yearly.

## I. New Developments

With the help of United Nations staff, models for developing countries were introduced into the LINK system at a very early stage. They included area models for Africa, Asia—South East and Pacific—Latin America, and the Middle East. These models emphasized trade, domestic accumulation of fixed capital, and internal spending. The UN staff also provided models for the centrally planned economies. From Wharton Econometrics, the project drew on the model of the Soviet Union that was developed independently of the LINK effort. Also, a model of the Peoples' Republic of China was developed specifically for LINK by Lawrence Lau of Stanford University, who was an early participant in the entire project.

\*Department of Economics, University of Pennsylvania, Philadelphia, PA 19104.



Over the years more countries were added to the LINK system. Now, virtually every OECD country is participating directly (with the exceptions of Iceland, Portugal, Turkey, Yugoslavia). To this expanded grouping of industrial countries, we are now able to include models for individual developing countries (with some small residual area groupings), which together with OECD and centrally planned economic (CPE) models, make up a system of 72 separate component models. These models are all interrelated through a commodity-specific trade matrix that is fully  $72 \times 72$ . This unusual statistical construct for trade has been prepared by the UN staff. At the earlier stage, the four area models of the developing countries were related to the rest of the system by only one row-column pair in the trade matrix. That limitation has now been broken, and the feedback interrelationships cover 72 entities at a completely bilateral level. It is, understandably, a mammoth job of information management and computing. The system now has about 16,000 simultaneous equations.

At the present level of disaggregation, there are still some unwanted groupings and the next stage of refinement will probably cover 80 separate models, with some major African countries being given separate treatment. To give a rough idea of the present scope of the system, the LINK project calculations produce some summary tables of key macroeconomic statistics for individual models, together with some broad groupings. These are taken from the present (December 1984) baseline projection, and are but an aggregative sample of the full system's results.

In order to facilitate the work of adding some 35 new models for developing countries on fairly short notice, we built similar (not identical) systems in Philadelphia, but several representatives from developing countries participate regularly in LINK, and we shall soon have fresh models maintained in a few specific developing countries as part of the system. These countries are Chile, Nigeria, Venezuela, Taiwan, South Korea, the Philippines, India, and Pakistan.

It is worthy of note that two models for centrally planned economies will be maintained on site, namely, Poland (University of Lodz) and Hungary (International Market

Research Institute). It is hoped that a model of the Peoples' Republic of China will be built domestically for use in LINK.

The inclusion of developing country models in LINK, with prominent display, has long been a distant goal, but with the emergence of specific problems for developing countries, their extended treatment became a priority of the first order. It has long been expected that developed countries would have a more modest economic growth prospect in the 1970's and in the years to come in the 1980's, than previously, especially the buoyant years of 1950-73. This indicates trouble for the developing countries, whose growth prospects depend heavily on performance in the developed countries. In addition, the appearance of the LDC debt problem, in bold relief after summer 1982, made their explicit study mandatory, in order to assess the chance of their being able to manage their economies without default or other crisis-precipitating events. There is significant feedback from developing to developed countries, and this is one of the objectives of investigation in the new enlarged system.

The models prepared for the developing countries have both a general and a specific group of structural equations. The general equations are those for various types of final demand, including exporting and importing. They also cover the generating of factor-income flows, prices, and wage rates. Specific aspects of these models vary from country to country. They show such things as

- 1) particular export markets for primary and other (service) products (coffee for Brazil, metals for Chile and Peru, grain for Argentina, overseas construction for Korea, Persian Gulf earnings for many poor countries, Suez revenues for Egypt, etc.);

- 2) sensitivity of money supply to external and to internal deficits; and

- 3) dependence of production on imports of raw materials and capital equipment.

When heavily indebted developing countries found themselves unable to service debt in the usual way, they cut back imports severely. This had adverse effects on the United States and other industrial countries. This feedback effect is one of the targets for research explanation in this phase of LINK

activity, but also the cutback in imports caused serious production declines in the developing countries that initiated the import reductions.

In most models for industrial countries, the import multipliers are negative. This is a standard Keynesian result, but for developing countries, their model structure should be different in that imports should vary *directly* with overall production. We saw extreme cases of this in Brazil and Mexico during 1982-83.

With these thoughts, models were specifically designed for developing countries that would deal effectively with such issues. At present, there are 35 LDC models in an extended LINK system. Scenario, multiplier, sensitivity and forecast work will bring out flaws in these models, and after a year or two, we shall be having a strong system on a very large scale of country coverage.

These LDC models were designed at LINK Central (University of Pennsylvania) using a great deal of insight and guidance from third-world scholars who have participated in LINK meetings for several years. They are from Nigeria, Venezuela, Chile, Taiwan, South Korea, India, and Pakistan. The project also received much help from staff economists of the UN and Asian Development Bank. Shinichi Ichimura of Kyoto University also contributed greatly to this development, on the basis of his studies of a sublinkage system in the Pacific Basin.

Gradually the focus of model responsibility will shift from LINK Central to particular economists in each of several developing countries. We are already installing models from some of the developing countries listed above. These models come from ongoing econometric practice in the particular countries, and their work will continue indefinitely into the future.

Related to the work on model building for developing countries in LINK, we are returning to an older theme, namely, the integration of commodity models with country models in the system. These enable us to study pricing and export earnings of primary producing countries many of whom are LDCs.

Additional studies of exchange rate determination, but not from the standpoint of

capital flows, were reported to the 1983 meeting of LINK at the Bank of Japan, Tokyo. Endogenous treatment of exchange rates will continue to be a lively area of study within LINK.

At the time of the freeing-up of exchange rates in the floating system, a problem was posed in LINK research as follows:

For fixed parities trade balances are endogenous, exchange rates are exogenous; and

For equilibrium rates under floating parities, trade balances (or current accounts) are set at exogenous target levels, exchange rates are estimated as endogenous instrument variables.

This makes exchange rate determination a problem of the well-known target and instrument formulation for economic policy formation in the manner of Tinbergen. This particular problem was formulated as one of equal numbers of instruments and targets. This attractive way of looking at the problem intrigued Keith Johnson of the LINK staff in 1974, and he studied the matter in his dissertation research. To solve this problem, he modified the LINK system into a much simpler model, with an identical master specification for each country. He solved the large-scale computational problem with the reduced system.

Christian Petersen and Peter Pauly of the present LINK staff reached a milestone in 1984 research by successfully programming the solution to this problem in a much expanded LINK system, ten years later. Their solution is very general and can be used for unequal numbers of targets and instruments as well as for the case of equality. The latter is strategic, however, because the loss function in the general case will be zero; therefore, this is a point that is straightforward for checking the enormously complicated computing problem. Petersen and Pauly were able to program the case of equality between numbers of instruments and targets for implementation of Ronald McKinnon's proposals for monetary stabilization by stabilizing exchange rates at purchasing power parity levels. These results were presented to a special LINK session held at the Federal Reserve Bank of San Francisco, August 1984, and will be reported in a volume of proceedings that is now being edited.

LINK researchers were able to solve the very difficult computing problem of optimal economic policy formation in the context of the multimodel system. In a broader sense, the whole computational apparatus of LINK has been much improved. It is more efficient, faster, cheaper, and more user friendly (with training). It is operated mainly on an interactive level from remote terminals now. Yet, it is still a lengthy calculation, taking about 20–30 minutes for a world simulation of about five-years horizon.

The new computer software for LINK was used to good advantage in a telecommunications experiment. The topic of the experiment was coordination of economic policy across country boundaries. Teams of LINK participants were asked to make live criticism of a basic LINK forecast, with a policy scenario of changed American choices between fiscal expansion and monetary restraint.

A team of European participants in LINK were assembled in a studio of British Telecom in London, facing a similar team of North Americans and a Japanese representative at SBS studios in McClean, Virginia. These two groups were in touch via satellite in an audio-visual mode. Two ("blind") groups with audio service alone were assembled at Los Angeles on the North American side, and at Geneva on the other side.

In a conventional audio-visual hookup, a general discussion can be held, looking at the other side's expressions while they hold forth. We added a new feature, namely audio-visual linkage together with a computer fully in the picture. After discussing the baseline forecast and the amendments, country-by-country, we interpreted for the model's computer inputs the quantitative meaning of the changes. The new information generated by the audio-visual exchange thus became the foundation for implementing a computer assisted response. It took nearly 20 minutes to execute a solution response to the changes in monetary and fiscal policies. The result appeared, finally, on computer screens that could be seen by all participants from both major teams in London and McClean. The computer screens with scenario information were also available to the participants in Geneva and in Los Angeles.

This calculation was something of a bottleneck, but we used the time fairly well in commenting about the whole projection, and when the final results were available on all screens simultaneously we were able to have a further teleconference discussion about the meaning of the findings.

LINK has shown how the power of the computer, working with a central data file, can be incorporated into an audio-visual exercise. This is what we wanted to show, in addition to analysis of the substantive issues at stake. We were able to show how government technicians, international civil servants, and interested academics can be in frequent touch for discussing projections or in working out new ones. The whole experiment lasted four hours.

The computing got completed for the telecommunications experiment, but it took a long time, using up much valuable satellite space for four hours. This suggests that new approaches to computing that could significantly reduce the time required for our standard scenarios, would greatly benefit the use of telecommunications methods. A goal should be the execution of a LINK response to verbal proposals in less than 5 minutes. Once that goal is met, it will be much easier and less expensive to use teleconferencing for coordinating economic policy.

New advances in computer method are entirely possible. The next step is to consider the *supercomputer*, which is obviously a hardware configuration that would deal with the inherent problems of simulating big systems, like the present augmented LINK model, and also serving to manage its large data files. But it is not bigness alone that suggests the supercomputer for the LINK problem. Speed is of the essence as indicated by the limitations of the computer associated teleconferencing exercise just discussed. The supercomputer does offer the possibility of significantly reducing the time required for lengthy simulations with the large LINK model for some 72 countries and areas. There is, however, one further aspect of natural fit between the LINK system and the architecture of the latest supercomputer. Some work by the principle of parallel processing, that is, multiple computer processing units are lined up, perhaps around a ring or lined up

along some other convenient configuration. The main point, however, is that the processors should be working simultaneously, doing computations all at once. The fit with the LINK system is that individual country models would not be solved purely serially, but would be solved simultaneously on the different processors, one per country. They would then all be sent to a linkage processor. In principle, this approach holds out great promise and is on our present research agenda.

Research analysis of the potentialities of the supercomputer is promising. It follows a "hardware" line of improvement. A "software" is also being explored through analysis of algorithmic efficiency. Work on a predecessor project of LINK, namely, the Brookings-SSRC model project for the U.S. economy fully established the Gauss-Seidel algorithm for simulating large scale systems. At the time (mid-1960's) large scale meant some 300 equations. Now we are dealing with thousands of simultaneous equations, and attention is being paid to the use of Newton's method for algorithmic efficiency. Newton's method was, in fact, investigated in connection with the development of simulation algorithms for the Brookings-SSRC model but abandoned in favor of the Gauss-Seidel method, which proved to be faster, cheaper, and easier to handle. The multi-model, in particular, appears to offer some natural advantages to Newton's method, and LINK computing research is now following that direction for improving the speed, cost, and efficiency of our present research work.

Related to the research investigations of the supercomputer, we are exploring another approach for parallel processing but not on the supercomputer hardware system. In collaboration with Noah Prywes of Pennsyl-

vania's Moore School of Electrical Engineering, LINK is exploring *cooperative processing*, that is, solving large-scale dynamic systems for simulation purposes, where the individual models are solved "at home" on whatever hardware is available with results being sent electronically to LINK central for consolidating consistently into a world model. People cooperating in highly dispersed centers represent a high degree of research collaboration and can probably achieve results that are similar to those produced by a supercomputer. A key factor in the use of cooperative computing is the ability to transmit data regularly, efficiently, and accurately to the central LINK files in Philadelphia. From our side, it would relieve a substantial amount of processing at LINK Central. With high-quality fellow workers, this aspect of the industry, in far-flung areas of the world, can be the highest form of collaborative effort in our subject.

## II. Conclusion

From the discussion of this paper, it can be seen that the system and practices of project LINK are far from being statically conceived. It is a dynamic and live project, exploring new ways for handling analyses of the developing countries, capital flows, exchange rates movements, teleconferencing with computers, and new ways to compute standard LINK tabulations more efficiently, more quickly, and more accurately.

## REFERENCE

- Klein, L. R., Pauly, Peter and Voisin, Pascal, "The World Economy—A Global Model," *Perspectives in Computing*, May 1982, 2, 4-17.

IN HONOR OF STEPHEN H. HYMER: THE FIRST QUARTER CENTURY  
OF THE THEORY OF FOREIGN DIRECT INVESTMENT†

The Influence of Hymer's Dissertation on the  
Theory of Foreign Direct Investment

By JOHN H. DUNNING AND ALAN M. RUGMAN\*

The great contribution of Stephen Hymer's seminal dissertation (1960) was to escape from the intellectual straightjacket of neo-classical-type trade and financial theory, and move us towards an analysis of the multinational enterprise (*MNE*) based upon industrial organization theory. The magnitude of this breakthrough can be put into perspective by considering the state of the art when Hymer wrote twenty-five years ago.

In 1960 the prevailing explanation of international capital movements relied exclusively upon a neoclassical financial theory of portfolio flows. In this frictionless world of perfect competition, with no transaction costs, capital moves in response to changes in interest rate (or profit) differentials (see Carl Iversen, 1936). According to this arbitrage theory, capital is assumed to be transacted between independent buyers and sellers, that is, there is no role for the *MNE*. At the time there was no separate theory of foreign direct investment (*FDI*). The work did not even ask the question, of "Why is there *FDI*?", despite the evidence of sectoral cross investments and the existence of large *MNEs* with intra-industry trade. If anything, the early work on *FDI* focused upon the "where" of investment in a particular nation or industry, for example, Dunning (1958) was chiefly interested in explaining U.S. *FDI*

in Britain. There was little interest in understanding the reasons for the *MNE*, or the nature of its operations.

The pioneering conceptual insight of Hymer was to break out of the arid mold of international trade and investment theory and focus attention upon the *MNE* per se. This permits us to treat *FDI* as a modality by which firms extend their territorial horizons abroad. The unique feature of *FDI* is a mechanism by which the *MNE* maintains control over productive activities outside its national boundaries, that is, *FDI* means international production. In this view, *FDI* is more than a process by which assets or claims are exchanged internationally (see Robert Aliber, 1970; 1983). Hymer's great insight was in focusing attention upon the *MNE* as the institution for international production, rather than international exchange.

Until Hymer articulated the process of *FDI* as an international extension of industrial organization theory, it was not possible to understand why the *MNE* transfers intermediate products such as knowledge or technology among its units across different nations while still retaining property rights over such assets. Today it is widely recognized that the theory of *FDI* (i.e., international production) is primarily about the transfer of nonfinancial and ownership-specific intangible assets by the *MNE*, which needs to appropriate and control the rate of use of its internalized advantage(s), see Rugman (1981), David Teece (1981; 1982), Richard Caves (1982) and Mark Casson (1983).

In this paper we first acknowledge Hymer's contribution to the theory of *FDI* and then move on to reinterpret his dissertation in the light of the modern theory of the *MNE*. We

†*Discussants*: Raymond Vernon, Harvard University; Robert Z. Aliber, University of Chicago; Paul Streeten, Boston University.

\*Professor of Economics, University of Reading, RG6 2AA England, and Director of the Centre for International Business Studies, Dalhousie University, Halifax Nova Scotia, B3H 1Z5 Canada, respectively.

find that Hymer's dissertation is remarkably prescient in its identification of structural market failure, but that it somewhat overlooks the transaction-cost side of the literature. This led Hymer to overemphasize the market-power advantages of *MNEs*, while his followers have misinterpreted the need for regulation of the *MNE* based on this predilection. We find that some parts of Hymer's dissertation provide useful linkages to the relatively new work on strategic management of *MNEs* in a world of global competitive rivalry.

### I. Structural or Transaction-Cost Market Imperfections?

The major value of Hymer's dissertation is its clear statement in chapters 1 and 2 of the industrial organization explanation of *FDI*. Here Hymer explains that the *MNE* is a creature of market imperfections. The *MNE* has the ability to use its international operations to separate markets and remove competition, or to exploit an advantage. Control over the use of assets transferred abroad is required by the *MNE* in order to minimize risks and to achieve monopolistic power. Hymer (1976) states that control of a foreign subsidiary "is desired in order to remove competition between that foreign enterprise and enterprises in other countries... or to appropriate fully the returns on certain skills and abilities" (p. 25). Later he states that the *MNE* "is a practical institutional device which substitutes for the market. The firm internalizes or supersedes the market" (p. 48).

Unfortunately, Hymer misses the distinction between structural and transaction-cost (cognitive) market imperfections made, for example, by Dunning (1981, p. 29). Hymer's entire analysis is based upon structural imperfections, which are Bain-type advantages to enhance the asset power of the *MNE*. They include scale economies, knowledge advantages, distribution networks, product diversification, and credit advantages. All of these help the *MNE* to close markets and thereby increase its market power. Hymer cites Joe Bain (1956) extensively for his analysis of structural market imperfections. On the other hand, cognitive imperfections are

Williamson-type transaction costs (see Oliver Williamson, 1975). These transaction costs arise naturally, or at least are assumed to be exogenous to the *MNE*. The *MNE* then responds to the transaction costs by creating an internal market. The latter type of market imperfections are inadequately treated by Hymer.

Once internalization has been achieved, the *MNE* experiences an ownership advantage which appears to be similar to a Bain-type asset power advantage. However, it is always important to distinguish the source of the advantage. If an exogenous market imperfection leads the *MNE* to organize an internal market as a substitute for either a missing regular (external) market, for example, in the pricing of knowledge, or to replace more expensive modes of transactions, then the process of internalization improves efficiency. No rents would be expected for the *MNE* when it is exercising transaction-cost power. In contrast, Hymer's concept of the *MNE* is restricted to the structural market imperfections view. He sees the *MNE* as having the power to close markets by using one or more of the Bain-type advantages. In turn, he believes that this leads to the generation of rents for the *MNE*.

It would seem to us that Hymer did not fully appreciate the work of Ronald Coase (1937) when he wrote in 1960. A close reading of Hymer's dissertation reveals an exclusive emphasis upon the structural market imperfections viewpoint, plus a discussion of the role of the tariff as an instrument inducing *FDI* instead of exporting as a mode of entry. No discussion of the Coasian theory of the firm can be found in Hymer. Despite the attempt of Charles Kindleberger (1984) to reinterpret Hymer's passing mention of information costs, it is clear that Hymer does not discuss them as transaction costs. It appears that Hymer did not embrace the full meaning and implications of market failure, and that his analysis, while not faulty, is limited to only part of the field of industrial organization. For a similar interpretation, see Teece (1981) and Casson in Rugman (1982).

In his dissertation, especially in chapter 3 in his discussion of *FDI* vs. licensing, Hymer does allude to "market impurities," which the casual reader might interpret as meaning

transaction costs. However, in these sections Hymer is again thinking of structural market imperfections, especially the alleged ability of *MNEs* to enjoy monopolistic benefits by closing markets. Hymer compares internalization to "contractual collusion," via cartels and other agreements between firms. He refers to Dunning's description of the tobacco industry in the early years of this century as an example of worldwide market sharing activities (p. 89). Yet at no point in the dissertation does Hymer explicitly or implicitly consider that hierarchical organizational structures can replace imperfect markets for reasons of efficiency. To Hymer, *MNEs* always exist for monopolistic reasons, that is, "to separate markets and prevent competition between units..." (p. 67).

Hymer also pays little attention to the location of *MNE* activity which is an important cog in the eclectic theory of international production (see Dunning, 1981). By this omission, Hymer neglects the importance of the geographical and spatial dimension of the *MNE* and the manner in which location-specific factors are determined interdependently with ownership- (firm-) specific factors in the process of *FDI*. The reader of Hymer does not get the message that internalization of transactions by the *MNE* is required in order to capture the externalities of separately related but commonly owned activities.

## II. Efficiency and Strategic Management

An interesting linkage of Hymer's asset power viewpoint can be made to the modern literature in strategic management. Michael Porter (1980) and others have suggested that the art of strategic management is to identify entry and exit barriers, and then operate generic strategies based upon price competition, product differentiation, or the seeking of market niches for product lines of the firm. Clearly these entry and exit barriers are virtually identical to the Bain-type advantages of which Hymer was so fond. What does this imply? Is Hymer the grandfather of strategic management as well as of the theory of the *MNE*? Probably not, but he should receive credit for directing attention towards

the ability of *MNEs* to close markets by the use of strategies which are now the basic tools of business management policy.

Hymer's view of the *MNE* as an institution exercising asset power spills over into some of his later, more radical, work where he writes about the collusive nature of monopoly capital (see the collection of his major papers in Hymer, 1979). This strand of his thinking is reflected in the dissertation, although much less strongly than in his later works. It was probably a result of Hymer's exposure to the U.S. economic environment of the 1950's, where there was a strong anti-trust literature. Since then the field of industrial organization has achieved greater maturity, and it is unlikely that Hymer's naive views of international oligopoly and collusive activity are consistent with modern thinking in the field, especially that summarized by Williamson (1981) in his discussion of bounded rationality, opportunism, and asset specificity.

As an example consider Hymer's treatment of vertical integration, which he perceives to be a monopoly device for providing extra profits. He misses the point of modern strategic management where vertical integration can be used as a competitive weapon against nonintegrated firms, to reduce costs or increase revenues. The great advantage of being an *MNE* is the ability to use internal markets across nations. The *MNE* can use transfer prices, maneuver liquid assets, move around production facilities, and so on. In this way the *MNE* has greater degrees of freedom than a uni-national firm confined to one country. Yet, as an offset, the *MNE* faces environmental uncertainty as foreign governments can change the political, cultural, and social factors which determine its economic efficiency.

The essence of this dispute about whether or not *MNEs* are efficient stems from the manner in which the *MNE* is modeled, especially the nature of the assumptions made about the endogeneity or exogeneity of the market imperfections. If the market imperfections are natural transaction costs, then the *MNE* is operating against nature and so internalization is conceptually efficient. However, if the market imperfections are

structural and endogenized, leading to asset power, then the *MNE* is best viewed as operating strategically, and such actions may or may not be efficient.

### III. Dynamics and Diversification

There are two other areas of theoretical interest in Hymer's dissertation. The first is his emphasis upon market structure and the dynamic nature of the ownership-specific advantages of *MNEs*. This is consistent with the dynamic modeling of Raymond Vernon (1966) and his doctoral students working upon product life cycles and the oligopolistic reaction of *MNEs*. Today, we recognize that the asset power of the *MNE* is threatened by rival *MNEs* who launch new product lines incorporating new technology or other types of firm-specific advantages. Again the compatibility of Hymer's work with the field of strategic management is acknowledged by this emphasis upon the changing nature and dynamic evolution of firm-specific advantages.

The second area of interest is Hymer's insight into the advantages of international diversification. He states "that profits in one country may be negatively correlated with profits of another country..." (p. 94), and that "an investor may be able to achieve greater stability in his profits by diversifying his portfolio and investing part in each country. This investment may be undertaken by shareholders of the firm, and not the firm itself..." (p. 95). In this section, Hymer foresees the later literature on the role of the *MNE* as an indirect vehicle for the achievement of the gains from international diversification in a world where individuals face transaction costs in undertaking such diversification themselves (see Rugman, 1979). This, in itself, was a considerable intuitive achievement of Hymer's, since twenty-five years ago the modern theory of finance was not yet developed. The mean-variance framework was not well understood and asset pricing models were not yet developed. It is ironic that the recent literature in international finance has ignored Hymer's correct insight in this area, while the literature on the regulation of *MNEs* has built upon his

somewhat flawed analysis of market imperfections.

### IV. Hymer and Policy

Hymer's dissertation does not deal with policy and he has little to say about the political or social issues of developing nations. He does not attempt to calculate the benefits and costs of *FDI* or technology transfer, nor does he analyze explicitly the impact of *MNEs* on developing nations. Despite his reputation as a radical economist, fully justified by his later work, Hymer's dissertation itself follows conventional lines. It is written broadly within the framework of mainstream economic principles of the time and it does not use *dependencia* arguments.

However, by his emphasis upon the market-closing ability of the *MNE*, Hymer shifted the emphasis from strictly efficiency-type evaluations of *FDI* towards the more sensitive issues of distribution of income and social justice. In this way Hymer's dissertation provides hints about the later development of his work in which he switched completely from the neoclassical efficiency paradigm to a radical critique of economics in which Marxist questions of distribution are at the forefront of his analysis.

Due to the theoretical impossibility of reconciling efficiency and equity arguments, a great debate has ensued over the last twenty-five years about the conflicting objectives of *MNEs* and nation states. Hymer's focus upon the ability of *MNEs* to limit markets and thereby increase their ability to earn rents helped to give an apparently plausible basis to this debate. His work has been interpreted by critics of the *MNEs* as a rationale for regulation of *MNEs*. This interpretation of Hymer's work is simplistic at best and, given his neglect of the transactional side of the theory of the *MNE*, is liable to be misleading.

What we have learned in the last twenty-five years is that the conflicts between the *MNEs* and nation states are infinitely more subtle than can be resolved by the partial nature of Hymer's dissertation work. These issues need to be analyzed today using models of disequilibrium dynamics. The older



static welfare criterion for evaluating the contribution of institutions, such as the MNE, to resource allocation has lost its appeal. In a disequilibrium situation it is not always apparent that the activities of MNEs lead to a more (or less) efficient allocation of resources. The answer depends upon an evaluation (often on a case-by-case basis) of their relative use of asset power in closing markets and of their ability to respond to transaction costs inefficiencies. To this unresolved debate the 1960 dissertation by Hymer still brings fresh insights, but fewer unambiguous policy implications than his followers have imagined.

#### REFERENCES

- Aliber, Robert Z., "A Theory of Direct Foreign Investment," in Charles P. Kindleberger, ed., *The International Corporation: A Symposium*, Cambridge: MIT Press, 1970.
- , "Money, Multinationals and Sovereigns," in Charles P. Kindleberger and David Audretsch, eds., *The Multinational Corporation in the 1980s*, Cambridge: MIT Press, 1983.
- Bain, Joe S., *Barriers to New Competition*, Cambridge: Harvard University Press, 1956.
- Casson, Mark, *The Growth of International Business*, London: Allen and Unwin, 1983.
- Caves, Richard E., *Multinational Enterprise and Economic Analysis*, Cambridge, New York: Cambridge University Press, 1982.
- Coase, Ronald H., "The Nature of the Firm," *Economica*, November 1937, 4, 386–405.
- Dunning, John H., *American Investment in British Manufacturing Industry*, London: Allen and Unwin, 1958.
- , *International Production and the Multinational Enterprise*, London: Allen and Unwin, 1981.
- Hymer, Stephen H., *The International Operations of National Firms: A Study of Direct Foreign Investment*, (1960), Cambridge: MIT Press, 1976.
- , Robert B. Cohen et al., eds., *The Multinational Corporation: A Radical Approach*, Cambridge: Cambridge University Press, 1979.
- Iversen, Carl, *International Capital Movements*, Oxford: Oxford University Press, 1936.
- Kindleberger, Charles P., *Multinational Excursions*, Cambridge: MIT Press, 1984.
- Porter, Michael E., *Competitive Strategy: Techniques for Analyzing Industries and Competitors*, New York: Free Press, 1980.
- Rugman, Alan M., *International Diversification and the Multinational Enterprise*, Lexington: D.C. Heath, 1979.
- , *Inside the Multinationals: The Economics of Internal Markets*, New York: Columbia University Press; London: Croom Helm, 1981.
- , *New Theories of the Multinational Enterprise*, New York: St. Martin's Press; London: Croom Helm, 1982.
- Teece, David J., "The Multinational Enterprise: Market Failure and Market Power Considerations," *Sloan Management Review*, September 1981, 22, 3–17.
- , "A Transaction Cost Theory of the Multinational Enterprise," Discussion Paper in International Investment and Business Studies No. 66, University of Reading, 1982.
- Vernon, Raymond, "International Investment and International Trade in the Product Cycle," *Quarterly Journal of Economics*, May 1966, 80, 190–207.
- Williamson, Oliver E., *Markets and Hierarchies: Analysis and Antitrust Implications: A Study of the Economics of Internal Organizations*, New York: Free Press, 1975.
- , "The Modern Corporation: Origin, Evolution, Attributes," *Journal of Economic Literature*, December 1981, 19, 1537–68.

# Multinational Enterprise, Internal Governance, and Industrial Organization

By DAVID J. TEECE\*

A multinational enterprise (*MNE*) is a firm that controls and manages production establishments located in at least two countries. In a Coasian sense, *MNEs* substitute for the market. They can be divided into two nonmutually exclusive types. One turns out essentially the same line of goods or services from each facility in several different locations, and will henceforth be referred to as horizontally integrated *MNEs*. The other produces outputs in some facilities which serve as inputs into other facilities located across national boundaries. These are vertically integrated *MNEs*.

Two conceptually distinct issues exist which need to be addressed in order to explain the *MNE*.<sup>1</sup> The first is the locational question of why production occurs where it does. The second is why certain production activities occur under the control of foreign enterprises while others do not. The locational issues are explained rather well by the standard theories of comparative costs. The control (internalization) issues cannot be explained by standard theory, yet these are central to a theory of the *MNE* as compared to a theory of international production. The former is concerned with explaining the nationality of the firms engaging in international production, while the latter simply explains the international distribution of the world's productive activities without concern for ownership and control patterns.

\*Professor of Business Administration, University of California, Berkeley, CA 94720. I thank Robert Aliber, Greg Hawkins, and Jim Wilcox for helpful comments.

<sup>1</sup>John Dunning (1981) develops a useful taxonomy which explains *MNEs* in terms of ownership (of rent-yielding assets), locational, and internalization advantages. For expositional reasons, I collapse what Dunning refers to as ownership factors into locational factors because it is the coupling of the two which explains where production ought to be located.

Until about two decades ago, economists viewed the *MNE* as simply an arbitrageur of capital, transferring equity capital from countries where returns were low to those where it was higher, earning the arbitrageurs rents and contributing to efficient resource allocation. The capital arbitrage theory of *MNE* predicted that *MNEs* would be headquartered in countries where the domestic marginal productivity of capital was relatively low, from which they will transfer capital to subsidiaries where the marginal productivity was higher.

As Stephen Hymer (1976) first observed, however, there are several features of direct foreign investment (*DFI*) and *MNE* which are inconsistent with this theory. The *MNEs* overwhelmingly finance their host-country operations in host-country capital markets. Furthermore, there are substantial cross flows of direct foreign investment (see Jean-Francois Hennart, 1982), as well as substantial concentration of *DFI* and *MNEs* in particular industries. These observations would be consistent with an arbitrage theory only if domestic capital markets were highly balkanized.

Despite the contrary evidence, Robert Aliber (1970, 1983) has reiterated a version of the capital arbitrage theory based on the identification of separable currency areas. The argument goes approximately as follows: there are substantial differences among countries in nominal and real interest rates. Because nominal interest rate differentials are poor forecasts of future changes in exchange rates, a wedge is introduced between returns on similar securities denominated in different currencies.

If portfolio managers had "perfectly" priced exchange risk, then the corporate managers would be reluctant to have their firms incur foreign exchange

risks...the parents of the subsidiaries take on the foreign exchange exposure because the corporate managers believe that the interest rate differentials are significantly larger than from the anticipated rate of change of the exchange rate...The expansion of firms across national borders is consistent with the view that corporate managers internalize the costs and risks of foreign exchange exposures at lower costs than portfolio investors.

[Aliber, 1983, p. 252]

The argument is flimsy and incomplete at best. No explanation of why corporate managers can internalize exchange rate risk at lower costs than portfolio managers is provided or even suggested. Moreover, as mentioned earlier, only a small part of direct foreign investment involves intercompany loans from parents to subsidiaries. Aliber claims otherwise but cites no evidence.<sup>2</sup>

In searching for a plausible theory of *DFI*, Hymer advanced two major tenets. One was that *DFI* was motivated by attempts to remove competition among enterprises in different countries, and the other was the *DFI* was motivated by domestic firms' attempts to increase the returns from the utilization of firms' special advantages (Hymer, 1976, p. 33). With respect to the second category, Hymer drew on Joe Bain (1956) to suggest that the source of the advantage could be in superior production techniques, imperfections in input markets which allow lower buying prices for established firms, and similar first-mover advantages. Possessing such special advantages, a national firm could be profitable outside the home country despite its relative ignorance of local conditions abroad. It would prefer to do so rather than

license in order to avoid possible bilateral monopoly situations, to enforce use restriction which could not be imposed contractually, to avoid the haggling between licensor and licensee associated with the evaluation of the value of the technology, and to better protect its advantage from the perils of misappropriation.

These were powerful insights, which laid the foundation for a completely new paradigm of the international firm. In one short thesis, Hymer transported the theory of direct foreign investment out of international trade and finance and into industrial organization. But the field of industrial organization circa 1960 did not have quite the richness it has today, and the library of concepts from which Hymer could borrow was especially sparse with respect to the economics of complex organizations.

Furthermore, Hymer was constrained by the absence of a clear welfare criteria for evaluating the *MNE*. In the absence of alternatives, Hymer seized upon perfect competition and Pareto optimality and inevitably arrived at troublesome conclusions. This is because he saw the *raison d'être* of the *MNE* as stemming from the "impurities of the market [that] would not arise in competitive industries" (1976, p. 86), which led him to the conclusion that "a restriction on direct investment or a policy to break up multinational corporations may be in some cases the only way of establishing a higher degree of competition in that industry...the underdeveloped countries need to devote an important share of their scarce resources to building up national enterprises..." (Hymer, 1970, p. 444). However, if the *MNE* is based on a special advantage, as Hymer claims, then the conclusion of this line of logic is that the special cost advantage should be abandoned in order to worship at the altar of perfect competition.

There is no doubt that Hymer's work represents a major contribution to the positive economics of the international firm. However, his thesis is misleading in its emphasis upon market power rather than efficiency, as is explained below. Relatedly, it does not provide a workable framework for assessing host-country control issues. The framework

<sup>2</sup>The Aliber theory may have some validity with respect to shifts in the total levels of foreign investment at the national level. Even here, "safe haven" and economic-growth considerations also appear to be important. The recent drop in direct foreign investment in the United States—\$7.0 billion in 1983, \$10.8 billion in 1982, 23.2 billion in 1981, 12.2 billion in 1980, and \$15.3 billion in 1979—certainly doesn't square well with the theory.

advanced is also static, and unable to grapple with intermediate organizational forms, such as cooperative and teaming agreements.

### **I. Monopoly vs. Efficiency Interpretations of the Multinational Enterprise**

As I have indicated elsewhere (1981, 1983), the essence of the multinational firm varies according to whether its investment abroad is primarily vertical or horizontal. The rationale for vertical direct investment stems from the efficient functioning of internal production and distribution systems when bilateral dependence with attendant strategic maneuvering and costly haggling might otherwise emerge (Oliver Williamson, 1975; myself, 1977, 1983; K. Monteverde and myself, 1982). By substituting an internal governance structure (i.e., vertical integration) capable of circumventing such problems, the multinational firm makes a substantial contribution to economic efficiency. Vertical integration permits specialized cost-saving equipment to be installed in both upstream and downstream locations with less risk that it will be idled by international disputes between enterprises of different nationality facing different incentives. Vertical direct foreign investment ought to be seen primarily as a response to market failure, and its explanation does not require appeal to classical market power considerations. Indeed, except under restrictive assumptions, vertical integration cannot be employed as a mechanism to extract monopoly or monopsony rents. Hymer avoids separate treatment of vertical direct foreign investment, but hints that sequential monopoly may be at its heart, blind to the spurious nature of leverage theories of vertical integration. In any event, Hymer appears unaware of the organizational efficiencies which can be associated with vertical integration, and misinterpreted the welfare implication of the *MNE* on this account.

The explanation of horizontal direct foreign investment requires the coupling of two requirements. One is that the firm possess a rent-yielding asset of some kind (for example, know-how) which warrants utilization in offshore production facilities, and the second

is that market transactions (know-how, licensing arrangements) are inferior to direct foreign investment as instruments for appropriating rents from the sale of the services of the asset in foreign markets. This is because of both the revenue-enhancing and cost-saving properties of the *MNE*. Direct foreign investment will be selected where contractual difficulties such as the ability to price know-how, or to write, execute, and enforce use restrictions governing technology transfer arrangements are anticipated. The intra-organizational mode for transferring technology is not only revenue enhancing but also cost saving, as empirical studies have demonstrated (my 1977 article). Hymer tacitly recognized these cost-saving aspects of horizontal direct foreign investment (1976, p. 50) and explicitly recognized its revenue-enhancing properties, which he chose to emphasize, bluntly stating that "monopoly problems pervade any discussion of international operations and direct investment" (1976, pp. 85–86). In a technical sense, Hymer was correct, in that in the presence of perfect competition and frictionless markets, the incentives for direct foreign investment disappear. However, applied welfare economists are unlikely to view any departure from perfect competition with alarm. What is needed, and what was not supplied by Hymer, was an assessment of the comparative welfare properties of multinational firms and the market alternatives. Hymer's comparisons of the *MNE* with the frictionless market fiction was inapposite and fueled host-country antagonism towards the *MNE*.

Hymer's thesis has stimulated a flurry of books and articles on the multinational firm, including P. J. Buckley and Mark Casson (1976) and Alan Rugman (1982). As Charles Kindleberger (1984, pp. 180–81) points out, there isn't all that much in this literature that cannot be found in Hymer. What is accomplished, however, is a reemphasis away from market power concerns and towards the economies associated with internalization, a literature which is derived from Williamson (1975). Departures from perfect competition in product markets and transactional advantages from internalization are needed to explain horizontal *DFI*, but Hymer chooses

only to emphasize and develop the former. The new literature therefore provides a much-needed balance.<sup>3</sup> The principal deficiency is that the internalization arguments are insufficiently micro analytic, failing to establish a structure which can discriminate between situations where direct foreign investment is more efficient than alternative organizational modes. Clearly, direct foreign investment isn't the most efficient mode of organization in all circumstances where a firm possesses an advantage which is the foundation of a business opportunity abroad. The contingent framework which is needed to create a theory of the *MNE* is sketched in my papers (1981, 1983), where it is shown that the incentives to use nonmarket modes for horizontal transactions swing upon the degree to which the technology in question cannot be evaluated, packaged, and protected using contractual instruments. In the case of vertical transactions, the presence of transaction-specific assets generating *ex post* "lock in" conditions is the key incentive driving foreign direct investment. Needless to say, these conditions don't exist all the time, and arm's length contracts will often suffice.

## II. Host-Country Controls

The transaction cost theory embedded in the above explanation of the *MNE* also contains some interesting policy conclusions. Besides maintaining that the *MNE* ought mainly to be regarded in economizing terms, the micro-analytic lens afforded by transaction-cost economics enables one to assess, from an efficiency perspective, the desirability of host countries exercising special controls over the activities of *MNEs*.

The need for special governance over the *MNE* depends in large measure on whether

the host country provides transaction-specific assets to support the *MNE*. This might be the case, for instance, if the host country dedicates roads and electric utilities to support a particular *MNE*, or if workers develop special skills to meet the needs of an *MNE* affiliate. When specialized assets are developed for the foreign firm, they cannot be redeployed easily should the *MNE* decide to pull out, or to otherwise adjust its plans opportunistically. The host country has incentives to develop investment safeguards in these circumstances. These can take many forms, including some type of penalty for early withdrawal or delayed expansion, or the creation of specialized governance structures for resolving disputes between the *MNE* and the host country. Alternatively, the *MNE* could be asked to post a bond, engaging in what Williamson (1983) refers to as the exchange of hostages. Despite safeguards of the type indicated, additional benefits might accrue from information disclosure involving relevant strategic plans, or informational representation on the Board of Directors, as this might enable the host country and workers with firm-specific skills to anticipate future developments and plan accordingly. Failure to design such governance structures may result in host countries failing to provide the necessary infrastructure. It bears repeating, however, that special governance or regulatory machinery is not needed unless the host country is deploying specialized assets to support the foreign firm.

The argument is also symmetrical and more empirically relevant when viewed from the *MNE's* perspective. If *MNEs* are to invest specialized assets dedicated to the host country, immobile plant and equipment being a case in point, investment safeguards are needed. An expropriation price agreed upon *ex ante* which the *MNE* can trigger at its discretion, somewhat analogous to certain types of severance pay in employment contracts, is one such possibility. Arbitration is another. Failure to design machinery of this kind will dissuade direct foreign investment when specialized nonredeployable assets are needed for efficient production. Alternatively, the *MNE* will extract a "fee" to offset this risk. The fee might take the form of

<sup>3</sup> It is noteworthy, however, that subsequent work in the spirit of Hymer, such as Richard Caves (1971), chose to emphasize the monopoly issues and neglected the internalization issues which are mentioned in Hymer but not well developed. Caves (1982) has subsequently redressed this misemphasis and signed on to the transaction cost thesis.

higher prices for goods sold in the host country, and/or lower prices for commodities purchased.

### III. Dynamic Considerations

The literature on the *MNE*, whether emphasizing market power or efficiency, suffers from a common deficiency: underemphasis on dynamics. The historical evidence shows American and British *MNEs* transitioning to offshore production after first establishing a sales branch abroad which in turn commonly preceded the establishment of contractual relationship with a foreign sales agent (Steven Nicholas, 1983). Agency contracts were unsatisfactory, containing vague and difficult to enforce performance clauses. Attempts to shore these up by specifying, quantitatively, budgets for traveling, advertising, engineering support, and the like, as well as certain inventory levels on the agents' premises proved unsatisfactory. The transition from agency to branch sales office was facilitated by the manufacturer's gradual accumulation of information about the foreign market, acquired through monitoring its foreign agents, and by expansion in sales volumes to levels which would support a facility of minimum economic size. The establishment of a sales branch also demonstrated to customers a more solid commitment on the part of the manufacturer to support the market in question. Often triggered by the failure or termination of an agent, the establishment of a foreign sales subsidiary subsequently became the platform upon which a manufacturing investment was made (Alfred Chandler, 1977, p. 369). However, the transition to manufacturing depended upon the relationship of production costs abroad to production costs at home plus tariffs and transportation, as well as governance considerations. Nevertheless, as Yair Aharoni (1966) points out, the direct investment process is governed by more than just economic incentives. An initiating force, galvanizing the organization into action, is often needed. The presence of a sales office also assists information collection, thereby significantly lowering perceptions of uncertainty, and raising the probability that a firm will engage

in direct foreign investment if the underlying cost conditions permit it. Clearly, transaction-cost economics must be married to organizational decision theory if the dynamics of channel selection are to be better understood.

### IV. Conclusion

The Hymer thesis cast the *MNE* and direct investment in a new light. The contribution was considerable, though the analysis was incomplete, and the welfare conclusions misleading. Considerable progress has been made in the last two decades in embellishing the theory and correcting the errors. More work needs to be done in understanding the distinctive characteristics of the *MNE*, its internal governance structure, and its incentive limits. Progress in this area is unlikely to be rapid until our understanding of the internal resource allocation and governance processes within firms begins to match our understanding of how these processes work in markets.

### REFERENCES

- Aharoni, Yair, *The Foreign Investment Decision Process*, Boston: Division of Research, Harvard Business School, 1966.
- Aliber, Robert Z., "Money, Multinationals and Sovereigns," in Charles P. Kindleberger and David B. Audretsch, eds., *The Multinational Corporation in the 1980s*, Cambridge: MIT Press, 1983, ch. 11.
- , "A Theory of Direct Foreign Investment," in C. P. Kindleberger, ed., *The International Corporation*, Cambridge: MIT Press, 1970, ch. 1.
- Bain, Joe S., *Barriers to New Competition*, Cambridge: Harvard University Press, 1956.
- Buckley, P. J. and Casson, M., *The Future of the Multinational Enterprise*, London: Holmes and Meier, 1976.
- Caves, R. E., "International Corporations: The Industrial Economics of Foreign Investment," *Economica*, February 1971, 38, 1-27.
- , *Multinational Enterprise and Economic Analysis*, Cambridge: Cambridge

- University Press, 1982.
- Chandler, Alfred**, *The Visible Hand*, Cambridge: Harvard University Press, 1977.
- Dunning, John**, *International Production and the Multinational Enterprise*, London: Allen and Unwin, 1981.
- Hennart, Jean-Francois**, *A Theory of Multinational Enterprise*, Ann Arbor: University of Michigan Press, 1982.
- Hymer, Stephen**, *The International Operations of National Firms: A Study of Direct Foreign Investment*, Cambridge: MIT Press, 1976.
- , "The Efficiency (Contradictions) of Multinational Corporations," *American Economic Review Proceedings*, May 1970, 60, 441-48.
- Kindleberger, C. P.**, *Multinational Excursions*, Cambridge: MIT Press, 1984.
- Monteverde, K. and Teece, D.**, "Supplier Switching Costs and Vertical Integration in the U.S. Automobile Industry," *Bell Journal of Economics*, Spring 1982, 13, 206-13.
- Nicholas, Steven**, "Agency Contracts, Institutional Modes, and the Transition to Foreign Direct Investment by British Manufacturing Multinationals Before 1939," *Journal of Economic History*, September 1983, 43, 675-86.
- Rugman, Alan M.**, *New Theories of Multinational Enterprises*, New York: St. Martins Press, 1982.
- Teece, David**, "Technological Transfer by Multinational Firms: The Resource Cost of International Technological Transfer," *Economic Journal*, June 1977, 87, 242-61.
- , "Multinational Enterprise: Market Failure and Market Power Considerations," *Sloan Management Review*, September 1981, 22, 3-17.
- , "Technological and Organizational Factors in the Theory of Multinational Enterprise," in Mark Casson, ed., *Growth of International Business*, London: Allen and Unwin, 1983, ch. 3.
- Williamson, Oliver E.**, *Markets and Hierarchies*, New York: Free Press, 1975.
- , "Credible Commitment: Using Hostages to Support Exchange," *American Economic Review*, September 1983, 73, 519-40.

# Hymer and Public Policy in *LDCs*

By DONALD J. LECRAW\*

Initially this paper was titled, "Stephen Hymer's Influence on Public Policy in *LDCs*." In an attempt to research this impossible subject, I conducted a citation search in the SSCI. The SSCI listed 442 citations in 192 different journals over the period 1977 to 1983, a count that places Hymer in the top five writers on international business and the multinational enterprise over that period. His writings were cited in journals devoted to law, history, philosophy, sociology, psychology, political science, geography, business, anthropology, broadcasting, peace research, statistics, migration, and urban, regional, agricultural, trade, financial, labor, industrial, and development economics; they were cited in journals devoted to the study of Africa, Asia, Latin America, as well as Europe, Australia, New Zealand, and North America; and they were cited by authors across the spectrum from arch conservative to radical political economists. The breadth of discipline, geography, and viewpoint of the authors who found Hymer's writings useful in their analysis reflects Hymer's multidisciplinary approach to problems, the originality of his insights, the clarity of his thought, and the forcefulness of his writing.

Although Hymer is best known for his work on the multinational enterprise (*MNE*), he also was intensely concerned with problems of less developed countries (*LDCs*) and the public policies they might use to alleviate the problems arising from trade and foreign direct investment (*FDI*). In his analyses, Hymer started from his base in economics, especially the economics of the *MNE*, but also used the insights of historical, political, and sociological analysis to illuminate these complex subjects. In using this approach Hymer may have fallen between two camps in

the analysis of public policy issues in *LDCs*: the "pure" economists (who have often viewed him as a bright, but misguided and ultimately fallen angel) on the one side, and the radical political economists (who have viewed him as a rising, if somewhat backward star) on the other side. Instead of bridging the gap, Hymer may have fallen into it: his insights cited, but his conclusions discounted as faulty or half-formed. This is unfortunate, since Hymer, writing in the 1960's and early 1970's, speaks directly to many of the concerns of today and of the future: the New International Economic Order, the North-South dialogue, and the basic needs of those at the bottom of the income distribution in high-income and low-income countries alike.

Hymer's writings addressed two major questions: how best should a small, developing country interact with the world economy through trade, inward (and outward) *FDI*, and technology licensing? How best should such a country organize its internal economic activity to meet the needs of all its people, especially those in the lower two-thirds of the income distribution?

To understand Hymer's approach to the analysis of these two questions and his contribution to public policy in *LDCs*, it is useful to set them within the context of his background, education and experience. Hymer was a Canadian. He grew up and received his first university degree in Canada, a country with a small, open economy, which largely exports raw materials and imports manufactured products, and whose manufacturing, energy, and mining sectors are dominated by subsidiaries of *MNEs*. In many ways, Canada was (and is) akin to the *LDCs* and his interests naturally turned in their direction. He was educated as an industrial organization economist at McGill and MIT. His experience and education, therefore, gave him a knowledge of large, multinational enterprises and the tools to analyze them. After MIT, Hymer worked in Ghana where he saw

\*Professor, School of Business Administration, University of Western Ontario, London, Ontario, N6A 3K7 Canada. I am grateful for partial funding from the Centre for International Business Studies and the Fund for Excellence, U.W.O.



the harsh effects of poverty in the aftermath of colonial imperialism and the neocolonial economic system of trade and the *MNE*. He soon concluded, however, that economics by itself was not a useful tool of analysis of the problems in *LDCs*—as he wrote: “Both in my study on Ghana and my study of direct foreign investment the disproportion between the questions I was asking and the tools I had to deal with them grew daily” (1979, p. 277).

Hymer returned to teach at Yale during the time of the “Greening of America,” when many were questioning both the value of the fruits of capitalist economic development and the costs of the political, social, and military system with which it was linked. In Canada, where Hymer worked with the Task Force on the Structure of the Canadian Economy in the late 1960’s, one of the forms of this revolt was a rise in economic nationalism which questioned the benefits and costs of Canada’s close links with (some called them the yoke of) the United States through trade, technology, and *MNEs*. The conclusions of this Task Force Report eventually led to the formation of FIRA (the Foreign Investment Review Agency) and Canada’s following the “Third Option” to attempt to diversify its trade away from the United States.

It is difficult to summarize and characterize Hymer’s conclusions for public policy in *LDCs*, since his method of analysis and conclusions changed and developed so substantially over his life. Yet it is also incorrect to concentrate on his later papers, since several of his early papers (see Hymer and Stephen Resnick, 1969) are among the most often cited in the literature. With this brief background, we can turn to the two fundamental questions for Hymer and public policy in *LDCs*.

### I. The Implications of an Open Economy for *LDCs*

Throughout his writings, Hymer recognized that international trade and *FDI* had the potential to increase world economic efficiency and accelerate the economic development of *LDCs* through the international market system.

The multinational corporations...will probably spread production more evenly over the world’s surface... [and] be a force for diffusing industrialization to the less developed countries....

[1971b, p. 119]

...[A]n industrial program which fails to open up the national market and indeed provides shoddy goods to substitute for goods previously imported from abroad can bring about a retreat [to inefficient production] and away from specialization and exchange. ...From the point of view of government policy, two important targets are a dynamic industrial sector...are a communications and transport system to facilitate the flow of goods, capital, and labor among all the trading partners....

[Hymer-Resnick, 1969, p. 505]

The question of income distribution both among and within countries was central to Hymer’s writings. While recognizing the potential benefits of trade and foreign direct investment, Hymer concluded that the institutional structure through which they were conducted led to a disproportionate share of the benefits being realized by high-income countries (and by small groups within those countries) and a disproportionate share of the costs being realized by low-income ones. His early industrial organization analysis of the *MNE* led him to the conclusion that firms engaged in *FDI* to preserve and extend their oligopoly power. In his thesis, Hymer (1976) concluded that in addition to the (small) efficiency loss in resource allocation arising from oligopolistic pricing by *MNEs* above the competitive price (arguably offset by the dynamic innovations of *MNEs*), the transfer rectangle from consumers to producers could accrue to stakeholders in the *MNE* located abroad. In addition, the *MNE* had the ability to reduce the host country’s participation in its supranormal profits through manipulation of transfer prices on international intrafirm transfers of goods, services, capital and technology. Similarly for the gains from international trade, Hymer and Resnick concluded: “Much of the

gains from export growth went to the government, to urban centers, and to local and foreign elites" (1970, p. 485).

Beyond these problems of the distribution of the benefits of an open economy, Hymer identified several other economic problems arising from international involvement: inappropriate products (which developed and catered to "wants" rather than to "needs"); inappropriate production technology; concentration and urbanization of economic activity; and labor force dislocation and the wasting of human capital (through "slash and burn" strategies of financial capital mobility).

In his early writings, Hymer raised the possibility that these economic costs to an open economy from international trade, capital, *FDI*, and licensing could be reduced by government regulation at the national and multinational level. This possibility was squarely in the tradition of Hymer's schooling of reformist government intervention. Ultimately, however, he concluded that these instruments of public policy would be largely ineffective against the power of *MNEs*, and the governments and segments of society that benefited from them and supported them. Such government intervention might increase the benefits flowing to *LDCs*, but not to the point at which the net economic, social and political benefits would be positive.

This negative outcome of international economic interaction was due to the unequal distribution of *power*—both economic power and political power—between the groups competing for the benefits generated by an open world economy. Hymer saw the *MNE* (and the international economy under the regime of the *MNE*) as organized in a complex pyramid of power, information, activity, and wealth. At the top of the pyramid were the head offices of *MNEs* located in a few metropolitan centers in high-income countries. From these heights, the *MNE* planned and controlled worldwide operations in *R&D*, production, and marketing. Below them in the hierarchy lay the coordinators located in regional centers in the home country and abroad. At the lowest level lay the specialized functional levels of production and distribution located in the countries of

the periphery. By this organizational structure, *MNEs* could reap the benefits of both decentralization and centralization, differentiation in production and marketing, and integration in coordination, planning and control. It also allowed *MNEs* to draw the surplus generated by their operations in trade and investment toward the top of the pyramid. At the same time it isolated the units (host-nation states and labor) at the bottom to render them powerless to seize the benefits of international exchange for themselves individually or to unite to achieve their goals as a group. "There is a close correspondence between the centralization of control within the corporation and the centralization of control within the international economy. The multinational corporation system thus does not seem to offer the world either national independence or equality" (Hymer, 1971c, pp. 8–9).

Similarly, Hymer concluded: "international trade... is often based on a division between superior and subordinate rather than a division between equals" (1971d, p. 12).

Even if the countries of the periphery could gain the information, expertise, determination, and power to deal with *MNEs* and to shift the balance of international trade toward their favor, the governments of the high-income countries, the homes of the *MNEs*, would intervene in their role as protectors and champions of the existing economic structure in general and *MNEs* in particular: "In the last analysis, markets come out of the barrel of a gun, and to establish an integrated world economy on capitalist lines requires the international mobilization of political power" (Hymer, 1972, p. 93).

Given these economic, organizational, and political realities, Hymer concluded that public policy in *LDCs* should be directed toward disengagement from the international system of trade and foreign direct investment both by blocking trade and *FDI*, and by heavily regulating them. By these actions, *LDCs* could develop to suit the basic needs of their people. Under this development path, *LDCs* "would have little need for multinational corporations" (Hymer, 1972, p. 105) and "some withdrawal from international

trade was necessary to make the life made possible by science pleasant and worthwhile" (Hymer, 1971d, p. 19).

## II. International Structure of Economic Activity

Given Hymer's public policy prescriptions that *LDCs* should distance themselves from the international economy, his second major question is how they should organize their internal economies to best accomplish this goal.

On this question, Hymer's views changed most radically and completely. In an early article (1971a), Hymer concluded that in Ghana both the British colonial administration and the Nkrumah socialist government impeded economic development and reduced future development potential by actively discriminating against the entrepreneurial class: the British with an aim of perpetuating empire and the post-independence government in order to channel economic activity through the state. Hymer concluded that in the future public policy would: "Hopefully not place much value on the development planning from *above* that has characterized most of Ghana's experience up until now" (1971a, p. 178).

Hymer soon changed his positive conclusions on the efficacy of using the entrepreneurial class in *LDCs* as the primary engine for development. He based this conclusion on three propositions. First, as described above, the international economic system in general and *MNEs* in particular are antithetical to development in *LDCs*. Second, the entrepreneurial (capitalist) class in *LDCs* are likely to be co-opted to join with *MNEs* in their exploitation of *LDCs* since "...in the last analysis their [the native capitalist class] interest lies with the international system" (1979, p. 259).

Third, even if the national capitalist class tried to act in the national interest, they "are very weak and in no way a match for international capital" (1979, p. 259) and would soon be bought out or forced out of business.

Since "national capital" cannot be relied upon to protect and foster national development, the best and, in fact, the only real alternative is "state capital":

The counterstrategy for the underdeveloped country is national planning. It seems to me that only such an organization could be efficient in mobilizing the strengths of the population as a whole in the interest of the population as a whole. [1979, p. 255]

The primary goal of such a development policy would be to provide minimum standards of health, education and food and clothing for the entire population. [1971c, p. 11]

Hymer (1979) concluded that national socialist planning would have several advantages: it would centralize national decision making at one point; allow scope for planning economies of scale; subordinate economic decisions to political and social ones; increase a nation's bargaining power, expertise, and information base in dealing with the international economy; create a polycentric system to facilitate international nation to nation interaction; and decrease the distinction between the periphery and the center of the world economy. Such a system, if implemented soon, might also be able to take advantage of the temporary breakdown in the *MNE* oligopolistic consensus and open conflicts between *MNEs* brought about by the emergence of European and Japanese *MNEs*.

Hymer recognizes that there might be economic costs to such a development policy, but: "the high productivity of the new technology allows countries greater scope for national independence, since it becomes far less urgent to economize on scarce resources" (Hymer and Robert Rowthorn, 1970, p. 87). There may also be political costs to such a system, but: "Problems of ensuring democracy within a region would remain, but they would seem to be more tractable than ones associated with world stratification by multinational corporations" (1979, p. 255).

## III. Summary

Hymer's writings on public policy in *LDCs* focused on one set of questions: "The question is not so much whether industry will grow rapidly, but who will organize it—na-

tional capital, state capital, or foreign capital" (1979, p. 250).

Hymer concluded that development via the use of socialist state capital was the only means by which LDCs could foster growth to benefit all their population since foreign capital and foreign trade with high-income countries lead to inequality and subjugation: "underdevelopment is created by the same process as development and forms, as it were, the ugly underbelly of affluence" (Hymer and Resnick, 1969-70, p. 191). To reverse this pattern of inequality it is necessary for LDCs: "...to build a system that will reflect as closely as possible what the heart of the people demands, we need a world economy where information can move freely between nations but capital, that is, power cannot" (Hymer, 1979, p. 255).

#### IV. Retrospective

Since so many of Hymer's writings were devoted to what will be and what should be in the long run, perhaps ten to fifteen years is not time enough for a fair assessment of the accuracy of his conclusions. Nonetheless, at least in the short run, it would be fair to say that, in general, things are not progressing along the path Hymer envisioned or desired.

The United States has disengaged in Vietnam and the major instances of overt imperialist aggression against LDCs are the victorious Vietnamese in Cambodia and Laos, and the Russians in Afghanistan. Trade unions, one possible counterbalance to the power of MNEs, are less of a political and economic force than ten years ago, with the exception of the Solidarity Trade Union in Poland, a socialist country. Following the example of Japan, LDCs led by Singapore, Hong Kong, Taiwan, and Korea have embarked on a more outward-looking trade and investment development strategy. Even China and India are beginning to look outward in trade and FDI. The oligopolistic consensus between MNEs from the United States, Europe, and Japan foretold by Hymer has not as yet materialized (in fact, competition has increased) despite increased interpenetration of home-country markets through FDI and

joint ventures. The MNEs and export-oriented national firms based in LDCs have grown in number, size, and power as LDCs have gained increased access to technology, capital and expertise in international business.

Yet, despite these developments, the LDCs as a group are now more, not less, dependent and subservient to high-income capitalist countries. Hymer correctly analyzed the current problems of LDC debt and trade protectionism:

...[T]hese countries [the LDCs] find that the major creditor countries compete vigorously in offering suppliers credit but form a united front when collecting debt. [1979, p. 247]

It is not clear that the West has the economic agents to...open up marketing channels for the underdeveloped world...or the ability to readjust its own structure of production to allow importation of goods from the periphery.

[Hymer-Resnick, 1969-70, p. 190]

It is unfortunate that Hymer is not here to help LDCs formulate public policy in response to these problems.

As a last ironic note, Canada, which so influenced Hymer's ideas, has turned away, at least for the present, from the path of nationalism and socialism he advocated: FIRA has been disbanded; the third option is dead; and free trade with the United States is being actively discussed by a new conservative government.

#### REFERENCES

- Hymer, Stephen, (1971a) "The Political Economy of the Gold Coast and Ghana," in G. Ranis, ed., *Government and Economic Development*, New Haven: Yale University Press, 1971.
- \_\_\_\_\_, (1971b) "The Multinational Corporation and the Law of Uneven Development," in J. W. Bhagwati, ed., *Economics and World Order*, New York: Macmillan, 1971.
- \_\_\_\_\_, (1971c) "Partners in Development:



The Multinational Corporation and its Allies," *Newstatements*, 1971, 1, 4-13.

\_\_\_\_\_, (1971d) "Robinson Crusoe and the Secret of Primitive Accumulation," *Monthly Review*, September 1971, 11-36.

\_\_\_\_\_, "The Internationalization of Capital," *Journal of Economic Issues*, March 1972, 6, 91-111.

\_\_\_\_\_, *The International Operations of National Firms: A Study of Direct Foreign Investment*, Cambridge: MIT Press, 1976.

\_\_\_\_\_, *The Multinational Corporation: A Radical Approach*, Cambridge: Cambridge University Press, 1979.

\_\_\_\_\_ and Resnick, Stephen, "The Crisis and Drama of the Global Partnership," *Inter-*

*national Journal*, Winter 1969-70, 25, 184-91.

\_\_\_\_\_ and \_\_\_\_\_, "A Model of an Agrarian Economy with Non-Agricultural Activities," *American Economic Review*, September 1969, 59, 493-506.

\_\_\_\_\_ and \_\_\_\_\_, "International Trade and Uneven Development," in J. W. Bhagwati et al., eds., *Trade, Balance of Payments and Growth*, Amsterdam: North-Holland, 1970.

\_\_\_\_\_ and Rowthorn, Robert, "Multinational Corporations and International Oligopoly: The Non-American Challenge," in Charles Kindleberger, ed., *The International Corporation*, Cambridge: MIT Press, 1970.

# HUMAN CAPITAL AND CULTURE: ANALYSES OF VARIATIONS IN LABOR MARKET PERFORMANCE<sup>†</sup>

## Religion and the Earnings Function

By NIGEL TOMES\*

Economics is fundamentally atheistic. Religious beliefs, practices, and behavior play no role in the life of homo economicus. Recently, however, there has been some interest in earnings differentials between religious groups. There are a number of motivations for this research. First, the analysis of earnings differentials by religion extends the analysis of intergroup differences beyond the narrow focus on black-white and male-female differences, to groups where differences in ability associated with religious, ethnic, or cultural factors may be of considerable importance. Second, religion may be an important dimension of family background and environment—family values, skills, endowments, goals and culture inherited or acquired during childhood—which influence earnings, the rate of return to human capital, and the intergenerational transmission of economic status.

This paper summarizes recent theoretical models and empirical research regarding the influence of religion on earnings and the rate of return to human capital. Empirical results from the 1981 Canadian Census are also reported.

### I. Theoretical Perspectives

The intergenerational transmission of religious affiliation appears substantial. Over 80 percent of adult U.S. males, belonging to a major religious group (Protestant/Catholic/Jewish)

claimed affiliation with the religion in which they were raised (see my 1984 article, Table 2). The institution of the family appears therefore to perform a central role in determining religious behavior.

Gary Becker and I (1979) provide a theoretical framework for analyzing the intergenerational transmission of economic status and other characteristics in which children are assumed to receive endowments determined by the reputation and "connections" of their families, the genetic constitutions of parents, and the learning, skills, and goals acquired through belonging to a particular family culture. Such endowments are potentially important determinants of human capital investments and the rate of return on such investments. Religious affiliation represents one measurable dimension of family endowments and culture.

More generally the formative years in the family can be viewed as producing not only human capital, but also "religious capital"—including ethical and moral codes of behavior governing consumption, the allocation of time and interpersonal relationships. Religious capital and human capital may interact, both in the accumulation process and in utilization. Greater ability or human capital may increase the capacity to learn and retain religious tenets. The role of Protestant Sunday School in the rise of literacy is consistent with such complementarity (Lee Soltow and Edward Stevens, 1981). Complementarity in production would lead families with higher desired stocks of religious capital to make greater investments in human capital, due to its additional benefits in the religious sphere. Other things equal, greater investments would result in lower returns to human capital.

Religious capital and human capital may also interact in the marketplace. Investing in

<sup>†</sup>*Discussants:* Barry Chiswick, University of Illinois-Chicago; Stephen Steinberg, Queens College, CUNY.

\*University of Western Ontario, London, Ontario N6A 5C2 Canada, and Economics Research Center/NORC.

religious capital may involve the acquisition of traits such as honesty, diligence, and reliability, that are compensated in the labor market and may affect the returns to human capital. Monitoring and sanctions by group members may ensure greater adherence to these tenets among members than by other individuals, creating incentives among employers to use religion as a signal for productive characteristics. Religious beliefs may prohibit or proscribe certain skills, occupations, and working practices (for example, the rejection by the Amish of modern technology). Such restrictions appear likely to lower earnings by reducing the choice set of religious adherents. Whether the returns to human capital are increased or reduced depends on such factors as whether prohibited occupations are human capital intensive.

One hypothesis regarding the interaction between religion and the return to human capital concerns the earnings of Jews. The argument, formalized by Reuven Brenner and Nicholas Kiefer (1981), is that because of their past cultural history of the expropriation of material wealth, Jews make greater investments in human capital, which is embodied and transportable, and receive a lower rate of return. In contrast, Becker has conjectured that the high incomes and achievements of Jews are explained by high marginal returns on human capital investments.

Religious affiliation that influences earnings and human capital investments should also affect other dimensions of family choice. In particular "quality-quantity" models emphasize that parental investments in children and fertility are jointly determined (see Becker's and my article, 1976). Thus high returns to the human capital of Jews may explain their low fertility levels. Conversely, religious tenets regarding fertility should also influence parental investments. Thus, if Roman Catholics face additional psychic costs of birth control, and this lowers the price of numbers of children, the resulting larger family size would tend to reduce investments in each child and raise the marginal returns on such investments.

Finally, there are other channels through which religion may influence earnings. Religious sanctions on divorce may increase the

expected duration of marriage and hence encourage greater specialization and division of labor between spouses raising the labor market skills and earnings of one spouse. Moreover, religious institutions may help resolve intergenerational conflicts regarding parental investments in children and reciprocal reimbursement and support of the elderly. However, so far, these possibilities have not been examined in the literature.

## II. Empirical Studies

Sociologists have devoted some attention to income differentials by religion. The finding that Jews receive higher incomes than non-Jews, controlling for a variety of characteristics, appears almost universal. The ranking of Catholics and Protestants has been more controversial. Andrew Greeley (1976), in particular, has argued forcefully that Catholics earn more than similar Protestants and are members of the American economic elite in terms of income. Other researchers report no significant difference. However, most of these studies analyze income rather than earnings and constrain the effects of religion to be additive, so that the returns to human capital are not permitted to vary across religious groups.

Recent studies by economists employ the human capital framework. Barry Chiswick (1983) examined the earnings of Jewish males using the 1970 U.S. Census. In the absence of information on religion, Chiswick identifies Jews as individuals with a Yiddish or Hebrew mother tongue, restricting his attention to second-generation Americans. Compared with non-Jews, Jewish men have 16 percent higher earnings (other things the same), a 20 percent higher rate of return to schooling (0.081 vs. 0.068) and a steeper experience-earnings profile—suggesting that American Jews are more productive in creating and using human capital. These results represent evidence against the hypothesis that the investment portfolio of Jews is "biased" towards human capital, since this implies a lower return to the human capital of Jews.

My article (1983) analyzes earnings differences by religion using the 1971 Canadian Census that contains information on reli-

gious affiliation and hence permits the direct identification of Jews. The estimated returns to Jews from schooling and experience exceed those of other non-Jews, and in a number of cases these differences are statistically significant. Given the small number of Jews in the sample, these findings buttress Chiswick's conclusion that Jews receive higher rate of return on human capital. By way of contrast, the payoff to Jews from a university degree, holding years of schooling constant (the "credential effect"), is less than to other groups, and does not differ significantly from zero.

Comparing other religious groups in Canada, the returns to Protestants from schooling, experience, and a university degree exceed those to Roman Catholics. The coefficients imply considerable differences across religious groups in the lifetime earnings (human wealth). Ronald Meng and Jim Sentance (1984) find that these differences persist when the usual measures of family background are included. The observation that Catholics make smaller investments than Protestants and receive a lower rate of return is inconsistent with the hypothesis that these differences are a reflection of divergent fertility behavior. Rather, differences in abilities, school quality, and other factors may result in lower returns to human capital to Catholic families.

My recent study (1984) analyzes the interaction of religion and earnings using U.S. survey data—the NORC General Social Survey. One advantage of these data is that they identify the religion in which an individual was raised—a measure that is clearly preferable in terms of exogeneity. When religious differences are constrained to be additive, apart from a Jewish differential, there is virtually no evidence that religious background affects earnings. This contrasts with Greeley's claims of a sizeable Catholic advantage, based on analysis of the same data. The explanation appears to lie in the use of family income rather than earnings and the limited set of regressors employed in Greeley's study. In separate earnings regressions the marginal returns to Catholics from college education exceed those to similar Protestants (see my article, 1984, Table 4).

This offsets the disadvantage to Catholics of lower precollege returns. Hence the largest Protestant-Catholic differential (about 18 percent) occurs amongst high school graduates who do not enter college. This differential narrows at higher levels of schooling and implies that the secular growth in education will narrow the earnings differential between major religious groups.

### III. Further Results

Tables 1 and 2 present additional results based on the 1981 Canadian Census. The sample consists of native-born males age 25–64 who worked in 1980. The empirical human capital specification is similar to that employed in my article (1983) (see also Notes to Table 1). In Table 1, the effects of religious affiliation are constrained to be additive—only the intercepts are permitted to differ across religious groups. The reference group is Protestants. There are two possible ways of identifying Jews—either by ethnic background or current religious affiliation. Not surprisingly these measures are highly correlated: 0.92. The coefficients reported in line 1 imply that, other things equal, Ethnic Jews earn 12.7 percent more than Protestants, while the residual group (no religion/other religion) earn 5.1 percent less than Protestants. Both these differences are highly significant. There is no Catholic-Protestant differential. If Jews are identified by religion (line 2), the Jewish advantage declines to 10.2 percent. When both Jewish ethnicity and religion are included, only the ethnic variable achieves significance (line 3).

The results reported in lines 1–3 may reflect the different spatial location of religious groups. Jews in particular are highly urbanized: 95 percent live in the 13 largest cities; 70 percent in the two largest cities: Toronto and Montreal. When the sample is restricted to the large cities (line 4), the Jewish differential declines to 7.25 percent—equivalent to the payoff from 1.8 years of schooling. The earnings disadvantage to the residual group increases slightly to 6 percent.

The final two lines of Table 1 divide the sample between Quebec and the rest of Canada, since Quebec differs linguistically,



TABLE 1—EARNINGS FUNCTIONS FOR RELIGIOUS GROUPS IN CANADA, 1980: ADDITIVE EFFECTS

Sample	Ethnic Jews	Religious Jews	Roman Catholics	No Religion /Other Religion
1) All Canada	0.120 (4.946)	—	0.004 (0.624)	−0.052 (6.787)
2) All Canada	—	0.097 (3.983)	0.005 (0.628)	−0.052 (6.813)
3) All Canada	0.155 (2.613)	−0.037 (0.640)	0.004 (0.593)	−0.052 (6.813)
4) 14 Cities	0.070 (2.943)	—	−0.010 (0.983)	−0.061 (5.868)
5) Canada excl. Quebec	0.107 (3.553)	—	−0.004 (0.522)	−0.051 (6.184)
6) Quebec	0.177 (4.110)	—	0.047 (2.032)	−0.070 (2.300)

Notes: Sample: Canadian-born males age 25–64 (inclusive), ethnic origin not Chinese or black, with positive weeks worked in 1980 and positive earnings (from employment plus self-employment), not in the armed forces and able to speak one of the official languages ( $N = 80,986$ : 1 percent Jews, 39 percent Protestants, and 47 percent Catholics).

Dependent variable: natural log of earnings (1980).

Other variables in the regression (coefficients not reported): Years of schooling, experience and its square, log of weeks worked, dummy variables for: language, provincial and urban location, ethnic origin, university degree, self-employment, government employment, and not married, intercept.

Coefficients are reported on dummy variables for religion. Reference group is Protestant: Anglican, Baptist, Lutheran, Mennonite, Pentecostal, Presbyterian or United Church (combined Methodist and Congregational).

Absolute value of  $t$ -statistics are shown in parentheses.

culturally, and perhaps economically from the rest of Canada. Earnings differentials by religion appear larger in Québec: Ethnic Jews earn 19.4 percent more, Catholics 4.8 percent more, and the residual group 6.8 percent less than Protestants.

Table 2 reports the coefficients on human capital variables when separate regressions are estimated for major religious groups. In terms of the returns to schooling there are two groups: Ethnic Jews and Protestants receive a rate of return of 4.4 percent (lines 1–2), while Catholics and the residual group receive returns of 3.4 percent (lines 3–4). However the schooling coefficient for Jews does not differ significantly from that of other groups. More generally, there is very

TABLE 2—EARNINGS FUNCTIONS FOR RELIGIOUS GROUPS: SELECTED COEFFICIENTS

Sample	Years of Schooling	Experience	University Degree
A. All of Canada			
1) Ethnic Jews	0.043 (3.485)	0.047 (5.800)	0.250 (2.920)
2) Protestants	0.044 (27.513)	0.037 (26.840)	0.158 (10.667)
3) Roman Catholics	0.034 (26.044)	0.034 (30.506)	0.224 (17.424)
4) No Religion/ Other Religion	0.034 (11.662)	0.039 (16.293)	0.166 (6.717)
B. 14 Cities			
5) Ethnic Jews	0.045 (3.549)	0.047 (5.512)	0.255 (2.885)
6) Protestants	0.049 (21.500)	0.045 (23.943)	0.157 (8.242)
7) Roman Catholics	0.035 (20.126)	0.039 (25.455)	0.233 (14.335)
8) No Religion/ Other Religion	0.037 (9.379)	0.043 (13.405)	0.164 (5.356)
C. Canada excl. Quebec			
9) Ethnic Jews	0.058 (3.798)	0.054 (5.327)	0.175 (1.663)
10) Protestants	0.043 (26.656)	0.036 (25.844)	0.159 (10.497)
11) Roman Catholics	0.034 (15.132)	0.030 (15.985)	0.154 (7.072)
12) No Religion/ Other Religion	0.034 (11.114)	0.039 (15.381)	0.155 (5.957)
D. Quebec			
13) Ethnic Jews	0.008 (0.387)	0.039 (2.700)	0.417 (2.836)
14) Protestants	0.059 (6.897)	0.048 (6.338)	0.148 (1.883)
15) Roman Catholics	0.033 (21.480)	0.037 (26.878)	0.272 (17.272)
16) No Religion/ Other Religion	0.038 (3.641)	0.051 (6.072)	0.255 (3.266)

Notes: See Table 1.

little evidence that the returns to Jews from human capital (schooling and experience) exceed those to other groups. The only instance of a significant Jewish differential is for Canada excluding Quebec where the experience coefficient for Jews exceeds that of Catholics (line 9 vs. 11).

Taken as a whole this is meager evidence that the returns to Jews from human capital exceed those of other groups. These Jewish-non-Jewish differentials are considerably smaller than those reported in my earlier article (1983) using comparable 1970 data.

This suggests that more recent Jewish cohorts are becoming assimilated into the mainstream of economic Canadian life. By 1980 possession of a university degree was as important for Jews as for other religious groups. This was not the case in 1970.

In contrast, the Catholic-Protestant differentials in returns to schooling and experience persisted through the 1970's. In each of the samples the rate of return to Protestants from schooling significantly exceeds that to Catholics. Moreover, the payoff to experience for Protestants exceeds that to Catholics in samples B and C. However, in contrast to the results for 1970, the payoffs to Catholics from a university degree in 1980 exceed those to Protestants. A closer examination reveals that this difference arises mainly in Quebec.

Because of the larger payoffs to Catholics from a university degree, the Protestant-Catholic differential does not change monotonically with the level of education. The difference in degree coefficients in lines 2 and 3 roughly compensates for the difference in schooling coefficients, so that four years of university with a degree yields the same earnings increment for both Protestants and Catholics. Put differently, the difference in the returns to schooling occurs mainly at the pre-university level.

Thus far empirical research on religion and earnings has focused exclusively on males. By ignoring women workers we neglect an increasingly important segment of the labor force and overlook the possibility that the sexual division of labor may differ significantly between religious groups. Preliminary results from the 1981 Canadian Census indicate that religious earnings differences among women are strikingly different than among men.<sup>1</sup> Using data on all working women, when coefficients are constrained to be equal, ethnic Jews earn 8.8 percent less ( $t = 2.77$ ) and Roman Catholics earn 3.4 percent more ( $t = 3.39$ ) than the Protestant reference group. Thus while Jewish males earn significantly *more*, their female

counterparts earn *less* than the Protestant reference group. We would expect these differentials to be associated with contrasting sexual divisions of labor with Jewish males more specialized and Catholic males less specialized in the labor market compared to Protestants. The converse patterns should apply to women. The pattern for the returns to schooling among women is also different: 0.026 ( $t = 1.245$ ) for ethnic Jews; 0.057 ( $t = 23.84$ ) for Protestants; and 0.061 ( $t = 30.68$ ) for Catholics. The returns to Jews do not exceed those of other groups, nor do the returns to Protestants exceed those to Catholics. In fact, none of the coefficient differences are significant.

#### IV. Conclusions

The returns to research by economists on religion and earnings have been small. One problem is that the lack of robust stylized facts leads to the rejection of most simple hypotheses. Earlier studies found that Jews receive higher returns to human capital, but not in recent Canadian data. In Canada, human capital returns to Protestants exceed those to Catholics in both 1970 and 1980 data, but this result does not generalize to the United States. One empirical generalization that seems to apply is that, other things equal, Jewish males earn more than other religious-ethnic groups. The results presented here suggest that the converse is true for females. Perhaps one hypothesis worth pursuing is that cultural and ethnic values lead to a much greater sexual division of labor in Jewish families. This may lead Jewish parents to invest differently in sons vs. daughters.

#### REFERENCES

- Becker, Gary S. and Tomes, Nigel, "An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility," *Journal of Political Economy*, December 1979, 87, 1153-89.
- \_\_\_\_\_, and \_\_\_\_\_, "Child Endowments and the Quantity and Quality of Children," *Journal of Political Economy*, August 1976, 84, S143-62.

<sup>1</sup> These results are not corrected for sample selection bias in labor force participation.

- Brenner, Reuven and Kiefer, Nicholas, "The Economics of Diaspora: Discrimination and Occupational Structure," *Economic Development and Cultural Change*, April 1981, 29, 517-33.
- Chiswick, Barry R., "The Earnings and Human Capital of American Jews," *Journal of Human Resources*, Summer 1983, 18, 313-35.
- Greeley, Andrew M., *Ethnicity, Denomination and Inequality*, Beverly Hills: Sage Publishers, 1976.
- Meng, Ronald and Sentance, Jim, "Religion and the Determination of Earnings: Further Results," *Canadian Journal of Economics*, August 1984, 17, 481-88.
- Soltow, Lee and Stevens, Edward, *The Rise of Literacy and the Common School in the United States*, Chicago: University of Chicago Press, 1981.
- Tomes, Nigel, "Religion and the Rate of Return on Human Capital: Evidence from Canada," *Canadian Journal of Economics*, February 1983, 16, 122-38.
- , "The Effects of Religion and Denomination on Earnings and the Returns to Human Capital," *Journal of Human Resources*, Fall 1984, 19, 472-88.

# Cultural Differences in Labor Force Participation Among Married Women

By CORDELIA W. REIMERS\*

Both the distribution of income and the role of ethnicity in economic behavior can be illuminated by an analysis of ethnic differences in married women's labor supply. Differences in wives' labor supply are the main source, beside differences in rates of female headship and wages, of the disparities in family income among racial and ethnic groups in the United States (see my 1984 article). Moreover, differences among ethnic subcultures may affect the labor supply of wives more than they influence many other types of economic behavior. Ethnic groups are distinguished by, among other things, views about male and female roles in the family and about wives and mothers working outside the home, as well as by the value placed on children, family size, household composition, and the education of women. These "cultural" differences may give rise to systematic differences in utility functions that lead to systematic differences in behavior by women in different ethnic or nativity groups who face the same constraints or opportunity set. Such cultural differences in utility functions no doubt are historically shaped by economic as well as other circumstances, and they evolve, but more slowly than the economic conditions. Ethnic differences in attitudes are, therefore, presumably more pronounced in the first generation of immigrants than in their American-born descendants.

These cultural attitudes may have both direct and indirect effects on wives' labor supply. They directly affect the allocation of time between home and market work by women with the *same* education, number of

children, etc. They also affect decisions about education, fertility, and other choices which in turn influence the market work opportunities and value of home time, and so indirectly affect labor force participation.

Virtually all of the numerous studies of black-white differences in female labor supply have found that black wives have higher labor force participation rates (*LFPR*) than whites, even after adjusting for differences in measured variables such as age, children, education, location, other family income, and wages. (For a summary of the results, see Mark Killingsworth, 1983, pp. 122, 195, 202, 404; and Phyllis Wallace, 1982, ch. 2.) Several explanations have been suggested—such as blacks' greater marital instability, their extended-family households, black husbands' lower wages and less stable employment—but none has proved satisfactory. It seems that black wives' higher labor force participation is in large part a cultural difference, rooted in the historical experience of blacks in America, and not explainable by current conditions alone.

No one has yet attempted to measure and account for the differences in wives' labor supply among the other ethnic and nativity groups, as this paper will do. These differences are large, with the variation in annual *LFPRs* among ethnic and nativity groups being greater than the variation in annual hours worked for those in the labor force, as shown in Table 1. The ranking of groups in terms of annual hours worked by those in the labor force differs from the ranking in terms of *LFPRs*. This suggests that different parameters govern the participation and hours worked decisions, perhaps because the groups face different fixed costs of working. These two aspects of labor supply therefore need to be analyzed separately. In this paper I focus on the differences in labor force participation rates.

\*Department of Economics, Hunter College, and Graduate School of the City University of New York, 695 Park Avenue, New York, NY 10021. I thank Cecilia Conrad for helpful discussions. A PSC-CUNY Research Award helped support this research.

TABLE 1—LABOR SUPPLY OF MARRIED WOMEN  
AGES 18–64

	Annual LFPR	Average Annual Hours: $H > 0$
U.S. Born		
Asians	.711	1639
Blacks	.664	1536
White Non-Hispanics	.589	1383
Hispanics	.553	1346
Foreign Born		
Asians	.536	1524
White Non-Hispanics	.533	1450
Hispanics	.467	1459

Source: 1976 *Survey of Income and Education* micro data file.

The American-born groups all have higher *LFPRs* than the foreign born. The U.S.-born Asian wives have the highest rate, followed by blacks, then American-born white non-Hispanics; while foreign-born Hispanic wives have the lowest rate. Part of these intergroup differences in married women's labor supply is no doubt due to differences in the opportunity set. Another part may be due to cultural differences in the utility function.

This paper is a first step toward explaining the differences in wives' labor supply among ethnic groups and between the native born and foreign born. I will ask to what extent married women's labor force participation varies across ethnic groups due to differences in their average age, health, location, and income levels; to what extent such culturally conditioned factors as language, family size and age structure, and education account for differences; and to what extent there is a cultural difference among ethnic groups in the extent to which women with the *same* language, number of children, and education work outside the home. Such an accounting will identify the questions that need further research to explain the intergroup differences in wives' labor force participation.

### I. Data and Methods

To answer the above questions, I estimate a reduced-form, linear probability model of labor force participation for married women age 18–64 for each of seven ethnic and nativity groups. I use data from the 1976 *Survey*

of *Income and Education* (U.S. Bureau of the Census). My model of labor force participation assumes that a woman is in the labor force if her expected market wage offer exceeds her reservation wage. The expected wage offer is determined by human capital characteristics and the reservation wage is determined by the value of home time and the fixed costs of working in the market. These in turn depend on personal and family characteristics.

My reduced-form equation has a dummy dependent variable indicating whether the woman was in the labor force (employed or unemployed) at least one week in 1975. On the right-hand side are a set of dummy and continuous variables, the determinants of the market wage offer and the reservation wage. The continuous variables include the husband's annual earnings in 1975; the family's other annual income in 1975 (excluding the wife's and husband's earnings, and earnings-related transfers); the AFDC benefit available if the family had no income; the number of children under age 6, ages 6–11, and ages 12–17; the number of other adults ages 18–64 and over age 65 in the household (besides the husband and wife); and the wife's educational attainment. Dummy variables indicate the wife's race and ethnic group, her age (in ten-year intervals), her date of immigration to the United States (if foreign born), whether she lacks fluent English, whether she had a non-English mother tongue (if U.S. born), whether she has a health limitation, whether she lives in a metropolitan area, whether her husband is over 65, and whether he is self-employed.

Having estimated this model, I then separate the difference in predicted *LFPRs* between U.S.-born white non-Hispanics and each ethnic-nativity group into several components: one set due to differences in average characteristics, and another set due to differences in parameters of the labor supply function. The results of this decomposition obviously depend on the weights applied to the differences in characteristics and parameters. My solution to this index-number problem is to use the average of the two groups' sets of weights. I interpret the difference in parameters as reflecting systematic cultural differences in the utility function across

ethnic-nativity groups, for women with given personal and family characteristics. Because many of the intergroup differences in average characteristics are themselves influenced by cultural attitudes (such as language, family size, education of women), I show separately the difference in *LFPRs* attributable to each type of characteristic.

The linear probability model is a linear approximation to a nonlinear probit or logit model, which makes it convenient for separating an intergroup difference in *LFPRs* into components. The estimated coefficients are nearly identical to the partial derivatives of the probit or logit model evaluated at the variables' means. Using the reduced-form estimates for this decomposition raises a potential problem: if, due to discrimination or labor market segmentation, members of a group receive a lower market wage offer than their human capital characteristics warrant and therefore participate less in the labor force, we would find a difference in the estimated parameters and would attribute their lower *LFPRs* to culture rather than to discrimination. However, this is not in fact a serious concern because there is little evidence of wage discrimination *among women* by race or ethnicity (see my 1985 article). In any case, in the present study I find no evidence of *lower* labor force participation due to a difference in parameters.

## II. Results

Tables 2 and 3 show the components of the differences in *LFPRs* between U.S.-born white non-Hispanics and other groups. (The means and estimated parameters of the labor force participation function for each group may be obtained from the author upon request.) According to these estimates, if the groups had the same average characteristics (other than ethnicity, race, date of immigration, and mother tongue), native-born white and foreign-born white and Hispanic wives would have virtually the same *LFPRs*, and U.S.-born Hispanics would have a 2 percentage-point *higher* rate than U.S.-born whites. Gone would be all of the gap between U.S.-born and foreign-born whites, and 95 percent of the gap between U.S.-born whites and foreign-born Hispanics. More-

TABLE 2—DECOMPOSITION OF DIFFERENCES IN *LFPRs* BETWEEN U.S.-BORN WHITE NON-HISPANIC WIVES AND FOREIGN-BORN WHITE AND HISPANIC WIVES

	Foreign-Born Whites	U.S.-Born Hispanics	Foreign-Born Hispanics
Total Difference	.056	.036	.121
Differences due to:			
<i>Characteristics</i> <sup>a</sup>	.057	.057	.115
Age (self, spouse)	.015	-.034	-.032
Health	-.0003	.001	-.005
Income	.005	-.018	-.020
Location (urban)	.007	.002	.002
Self-Employed Husband	.001	.003	.003
Family Size and Ages	-.007	.045	.051
Education	.024	.057	.054
Current English	.012	.0005	.062
<i>Intercepts and Slopes</i>	-.001	.022	.004
Intercept, including			
Ethnicity, Race <sup>b</sup>	-.038	.088	.044
Entry Date pre-1973 <sup>c</sup>	-.133	—	-.140
Mother Tongue	—	-.027	—
Other Slopes	.170	-.083	.104

Source: See text. Tables of group means and coefficients are available from the author upon request.

<sup>a</sup>Except ethnicity, date of entry, and mother tongue. The average of the two groups' weights are used. Totals may not agree due to rounding.

<sup>b</sup>For differences with foreign born, this also includes the effect of a 1973–76 entry date and the effect of mother tongue for U.S.-born whites (variable omitted from equations for foreign born).

<sup>c</sup>Not included in equations for U.S. born.

TABLE 3—DECOMPOSITION OF DIFFERENCES IN *LFPRs* BETWEEN U.S.-BORN WHITE NON-HISPANIC WIVES, AND ASIAN AND BLACK WIVES

	U.S.-Born Asians	Foreign-Born Asians	U.S.-Born Blacks
Total Difference	-.122	.052	-.076
Differences due to:			
<i>Characteristics</i> <sup>a</sup>	-.007	.040	.042
Age (self, spouse)	.003	-.037	-.005
Health	-.009	-.010	.013
Income	.020	-.007	-.019
Location (urban)	.007	.004	.004
Self-Employed			
Husband	.002	.002	.005
Family Size and Ages	-.011	.033	.013
Education	-.002	.012	.031
Current English	-.003	.043	—
<i>Intercepts and Slopes</i>	-.130	.013	-.118
Intercept, including			
Ethnicity, Race <sup>b</sup>	.224	.117	.117
Entry Date pre-1973 <sup>c</sup>	—	-.067	—
Mother Tongue	-.009	—	.004
Other Slopes	-.345	-.037	-.239

Source and Footnotes: See Table 2.

over, 75 percent of the gap between U.S.-born whites and foreign-born Asians would be closed, leaving only a 1.3 percentage-point difference in *LFPRs*.

On the other hand, the gaps between U.S.-born whites and U.S.-born Asians and blacks would be *widened* (substantially so for blacks), so that black and Asian American wives would both have *LFPRs* 12 to 13 percentage points above U.S.-born whites. Thus, differences in characteristics fully account for the lower *LFPRs* of Hispanic and foreign-born white and Asian wives, but completely fail to explain the higher *LFPRs* of U.S.-born Asians and blacks. Of course, cultural differences may be lurking behind some of the differences in characteristics, especially language, family size and age structure, and education. I next examine the effects of particular characteristics on wives' *LFPRs* and their contribution to the difference among ethnic and nativity groups.

As immigrant groups gradually assimilate to American society, cultural factors may result in labor supply differences between second generation women and those from the third and later generations in the United States. Having a non-English mother tongue (i.e., language used in one's childhood home) may therefore help explain labor supply differences among the U.S. born of different ethnic groups. I therefore include the effects of mother tongue among the differences in parameters that I attribute to "culture." However, this variable, which always has a positive coefficient, does not contribute to explaining the labor force participation gap for any group.

Lack of fluency in English is the most important difference between foreign-born Hispanics and Asians and U.S.-born whites. It explains a *LFPR* difference of .043 for foreign-born Asian wives, over 80 percent of their gap with U.S.-born whites. It explains a gap of .06 for foreign-born Hispanics, half of the total difference. For foreign-born whites, English explains a .01 difference, or only about 20 percent of the gap.

Differences in family size and age structure account for about a .05 gap in *LFPRs* between U.S.-born whites and Hispanics and a .03 gap for foreign-born Asians, whose

larger number of other adults partially offsets their larger number of children. This is more than the total difference between U.S.-born whites and U.S.-born Hispanics and a little over half of the difference for foreign-born Asians; but for foreign-born Hispanics, it is less than half of the *LFPR* gap. Blacks' larger families tend to reduce their labor force participation, so this cannot help explain their higher *LFPR*. The U.S.-born Asians have the same number of children as whites, but have more other adults and elderly at home. Both of these increase their *LFPR* (though the estimate is not significant), but this only explains 1 percentage point of the 12-point gap between them and U.S.-born whites. The foreign-born and U.S.-born whites are so similar in family size and structure that this cannot account for any of the *LFPR* gap between them.

Differences in education are among the biggest factors in the *LFPR* differences. Education explains a gap of nearly 6 percentage points between U.S.-born whites and U.S.-born Hispanics (more than 100 percent of the difference). It accounts for gaps of 5.4 points for foreign-born Hispanics, 2.4 points for foreign-born whites, and 1.2 points for foreign-born Asians. These are less than half the differences in *LFPRs* for these groups. Since U.S.-born Asian and white wives have the same education levels, this cannot account for any difference between them in labor supply. For blacks, the education difference, like the difference in number of children, works to lower labor force participation. Neither can help explain the black wives' high *LFPRs*.

The variables for immigrants' entry date and for ethnicity act as intercept-shifters, and their effects are attributed to cultural differences between groups. The entry date may also reflect an aspect of human capital: the acquisition of skills and information specific to the U.S. labor market. Insofar as I have wrongly attributed a human capital difference to a difference in parameters, my estimates of the difference due to culture are positively biased. If so, a more accurate measure would show the foreign born having even higher participation rates, if their characteristics were the same as U.S.-born whites.

My estimates provide strong evidence that immigrant women take just a few years to adjust to the United States before entering the labor force. In all groups, those who have been here at least four years are significantly more likely to be in the labor force than new arrivees. Labor force participation does not appear to change much with time in the United States after the first three years, aside from special "entry cohort" effects, such as the 1960-64 Cuban or the pre-1965 Asian immigrants.

My model also includes sets of dummy variables identifying ethnic origin insofar as possible. These dummy variables show any remaining difference in average *LFPRs* between ethnic groups, after all the measured characteristics have been taken into account. They therefore reflect what I interpret as a cultural difference in labor supply. For example, wives of German and British origin have the highest *LFPRs*, and Italians have the lowest among the identifiable groups of U.S.-born whites, given the same measured characteristics. Cuban wives have the highest and Puerto Ricans the lowest *LFPRs* among Hispanics of both nativity groups. Japanese American women have the highest *LFPR* among the U.S.-born Asian groups, but Filipino women have the highest rate among foreign-born Asians. Since 50 percent of the U.S.-born Asian wives are Japanese American, their high *LFPR* explains almost half of the Asian Americans' 12-point higher rate.

### III. Summary and Conclusions

I conclude, then, that the lower labor force participation of foreign-born white and U.S.-born and foreign-born Hispanic wives than U.S.-born whites is entirely due to differences in measured characteristics; no cultural difference in parameters appears to exist. A difference in parameters contributes a small amount to explaining the gap between U.S.-born white and foreign-born Asian wives; but it, too, is due mainly to the difference in characteristics—especially culturally influenced ones: language, family size and age structure, and education. Given the popular image of Hispanic culture, my finding that Hispanic wives who have the same schooling,

English, and family size as Anglo women participate even more in the labor force may come as surprise, but Vilma Ortiz and Rosemary Cooney (1984) also find traditional beliefs per se are unimportant for behavior. The influence of the Hispanic culture on married women's labor force participation works entirely through fertility, education, and language.

In contrast, for U.S.-born Asian and black wives, the entire difference from U.S.-born whites is due to direct cultural effects on the parameters of the labor force participation function, which account for a 12 to 13 point gap. The U.S.-born Asians (half of whom are of Japanese ancestry) have characteristics quite similar to U.S.-born whites. For blacks the cultural difference is larger than the actual difference in labor force participation; their characteristics lower their *LFPR*. Current personal and family circumstances cannot explain the high *LFPRs* for U.S.-born Asian and black married women; apparently we must turn to history to understand the cultural patterns that produce them.

### REFERENCES

- Killingsworth, Mark R., *Labor Supply*, Cambridge: Cambridge University Press, 1983.
- Ortiz, Vilma and Cooney, Rosemary S., "Sex-Role Attitudes and Labor Force Participation among Young Hispanic Females and Non-Hispanic White Females," *Social Science Quarterly*, June 1984, 65, 392-400.
- Reimers, Cordelia W., "Sources of the Family Income Differentials among Hispanics, Blacks, and White Non-Hispanics," *American Journal of Sociology*, January 1984, 89, 889-903.
- \_\_\_\_\_, "A Comparative Analysis of the Wages of Hispanics, Blacks, and Non-Hispanic Whites," in G. Borjas and M. Tienda, eds., *Hispanics in the U.S. Economy*, New York: Academic Press, 1985.
- Wallace, Phyllis, *Black Women in the Labor Force*, Cambridge: MIT Press, 1982.
- U.S. Department of Commerce, Bureau of the Census, *Microdata from the Survey of Income and Education*, Data Access Description No. 42, Washington: USGPO, January 1978.



# Peddlers Forever?: Culture, Competition, and Discrimination

By WILLIAM A. DARITY, JR. AND RHONDA M. WILLIAMS\*

Contemporary economic theory has all but completed the burial of the idea that market discrimination explains racial wage differentials or differences in general pecuniary accomplishments across ethnic groups under competitive conditions. We need only await the eulogy.

Interment began with the failure of the model premised on employers' taste for discrimination. The preferences of employers for members of one group over another could not sustain wage differentials under competition if the individuals from each group were equally able. Although the initial argument was made under the assumptions of neoclassical perfect competition, it was beaten back by the Austrian process view of competition (see Israel Kirzner, 1973). A latent reservoir of alert entrepreneurs presumably would seize the profit opportunities generated by the discriminatory wage gap, drive discriminating employers from the market, and erode the wage differentials.

Valiant efforts emerged subsequently to raise from the dead the idea that competition might eliminate market discrimination. These efforts involved the development of market discrimination models under the states of affairs characteristic of neoclassical imperfect competition or so-called "statistical" discrimination. Neither case precluded the possibility of ingenious entrepreneurship nor "the entrepreneurial capacity to smell profits" (Kirzner, p. 229). As a result, in neither case could the existence of wage or earnings differentials be maintained by employer decisions. In the first case, the inventive en-

trepreneur could circumvent or bring down the barrier that was the source of the imperfection. Similarly, in the second case the clever entrepreneur could devise new procedures for overcoming existing informational discrepancies that might exist about the abilities of the members of each of the ascriptively distinct groups (see Darity, 1982). The application of the Austrian process view of competition to the problem of racial and ethnic wage differences stripped market discrimination of its analytic significance.

However, the empirical persistence of such wage differentials in the U.S. economy is well established. The demise of the theory of market discrimination under competitive conditions has led to increasing use of the human capital explanation for differences in economic achievement across ascriptively distinct groups. After many years and scores of statistical studies that revealed that convergence in observable human capital characteristics—especially years of formal education—did not lead to the anticipated abolition of black-white wage gaps (see Williams, 1984), there is now a shift underway that identifies the source of the alleged human capital differences as unobservable cultural differences between ascriptively distinct groups. Cultural variation is accorded primacy in explaining ethnic and racial differences in economic achievement. In light of the death of the market discrimination explanation, the culturalological explanation is given added force by the observation that certain ethnic groups (Japanese and Jewish Americans, to name but two) have managed to succeed despite discrimination (see Barry Chiswick, 1983a, b).

The themes of culture and competition thus are the focus of this essay. We first provide substantive and historical critiques of the "new" cultural variant of the human capital theory. Second, we argue that the existence of ongoing market discrimination can be revived by employing an alternative

\*University of North Carolina, Chapel Hill, NC 27514, and University of Texas, Austin, TX 78712, respectively. We are grateful to Art Goldsmith, Bobbie Horn, Steve Steib, and David Swinton for helpful comments. Research support was provided by the Southern Center for Public Policy Studies at Clark College in Atlanta.

conception of competition rather than partaking in the wholesale denial of the importance of market discrimination. Specifically, classical and Marxist competition are proposed as alternatives to the neoclassical and Austrian views of competition. The former can be reconciled with the existence of market discrimination as a persistent source of differences in racial and ethnic group achievements.

### I. Culture as Human Capital: Logical Foundations?

Both Thomas Sowell (1981a, b; 1983; 1984) and Chiswick (1983a, b) have been in the vanguard of those proposing cultural explanations of racial and ethnic success or failure in the marketplace. Sowell describes culture as "ultimately ways of accomplishing things" (1983, p. 136) and argues that culturally produced attitudes and work habits often are the crucial determinants of group achievement. Chiswick conjectures that the higher rate of return to education enjoyed by second-generation Jewish immigrants to the United States: "may arise from cultural characteristics that enable Jews to acquire more units of human capital per dollar of investment...or it may be that there are cultural characteristics that enable Jews to be more productive in the labor market with the human capital embodied in them" (1983b, p. 334).

Moreover, Sowell in particular suggests that cultural differences are longstanding. In a telling discussion of the persistence of cultural inheritance, Sowell refers to the Jewish peddlers who trailed the Roman Empire's armies and their counterparts 2000 years later:

...[T]he reality of group patterns that transcend any given society cannot be denied. Jewish peddlers followed in the wake of the Roman legions and sold goods in the conquered territories. How surprising is it to find Jewish peddlers on the American frontier or on the sidewalks of New York 2000 years later—or in many other places in between? [1984, pp. 28–29]

Thus Sowell invokes the persistence of cultural traits and the continued possession of that culture by the direct descendents of the persons with those traits. But why do these traits persist, and why do they remain contained within a particular group? Sowell dismisses biological determinism in favor of pure cultural determinism with the cavalier remark that "no one needs to believe that Jews are genetically peddlers" (1984, p. 29). At the turn of the century, however, when intellectual conflict was greatest among proponents of culturally and biologically determinist views of human social evolution, there were those among the latter group who claimed the phenomenon of culture was "subordinate to nature" (see Derek Freeman, p. 21). The cultural determinists tended to argue that culture was self-organizing and self-reproducing. In what follows, we consider the logic and implications of cultural determinism.

From the pure cultural determinist perspective, culture is an attribute that can be changed, albeit with difficulty. Among economists, Richard Easterlin (1981) has advanced the view that (i) culture is malleable, and (ii) the whole world will become developed and culturally uniform—from the standpoint of the traits pertinent to economic advancement—as modern education simultaneously transmits those traits and signals their spread. As long as the predisposition toward certain cultural traits is not treated as innate or inherent—as given in nature—it can be transferred between members of various groups.

Furthermore, if some groups possess cultural attributes that enhance their performance in an environment where market exchange is a major aspect of social life, then we would expect incentives to emerge for members of the culturally advantaged group to transfer those attributes to less favorably endowed groups. The "new" cultural variant of the human capital theory presumes a persistence of market-valued cultural differences that is at odds with the conception of competition that undermined the theoretical persistence of market discrimination.

If ethnic differences in economic achievement are long-lasting because cultural dif-

ferences are long-lasting, then logically there must be a failure of the market mechanism to complete the transfer of the appropriate cultural norms from the high achievers to the low achievers. Hence, the variation in ethnic or racial economic accomplishment has to be explained by a failure of competition, which must mean a failure of entrepreneurs to recognize the source of variations and conduct the transfer. But why should the alert and inventive entrepreneurs fail here when they are assumed to have succeeded so admirably in destroying market discrimination? The Austrian competitive process that ostensibly can destroy a host of impediments to the free operation of the market is somehow immobilized when confronted with cultural differentiation. It was precisely because the competitive process was seen as so strongly operative that the cultural argument has gained such sway in some quarters of the profession, in spite of the logical inconsistency. Either Austrian entrepreneurs can undercut both market discrimination and cultural differentiation, or they can do neither.

One would expect cultural differences to persist if the more successful ethnic group can withhold its "trade secrets" from the marketplace. This means the economic theorist must explain how the successful group enforces sanctions against cultural transfers in the presence of opposing incentives the market might offer—that is, how do they prevent cartel cheating on an indefinite basis?

To recapitulate: if one accepts the position that cultural differences are significant determinants of ethnic and racial economic inequality, then one must explain why cultural differences salient for success are persistent. This persistence is impossible to maintain if one believes the market system is imbued with an Austrian process of competition. Sowell (1981b), for one, seems to have in mind exactly such a view of competition.

## II. Culture as Human Capital: Historical Foundation?

After comparing the socioeconomic achievements of various Asian-American ethnic groups (the Chinese, the Japanese, and the Filipinos), Chiswick notes that: "The

findings for the Chinese and Japanese suggest also that it is incorrect to assume that racial minority status in the United States and racial discrimination *per se* result in lower observed levels of earnings, schooling, employment, and rates of return to schooling" (1983a, p. 212). Chiswick correctly calls our attention to the substance of discrimination and leads us to inquire whether all racially identifiable nonwhite groups experienced similar patterns of "discrimination *per se*."

Considering first the case of American Jews, it has by now been well documented that Jewish immigrants from Eastern Europe made major income and occupational gains well before their children attended, en masse, those public institutions of higher education which provided avenues for the upward mobility of subsequent generations (see Selma Berrol, 1976; Sowell, 1981a, pp. 90–91; Stephen Steinberg, 1981, chs. 3, 5). The rapid socioeconomic progress of the immigrants was, in turn, facilitated by the historical juxtaposition of their unique constellation of skills and a growing, dynamic economy very much in need of those talents. Excluded from land ownership in Eastern Europe, urban Jews developed the petit bourgeois skills (including a high degree of literacy) which served them so well once entrenched in the urban centers of turn-of-the-century industrial American capitalism. Moreover, the entrepreneurial beginnings of Jewish peddlers, merchants, and clothiers provided the material base for the educational and occupational mobility of the next generation of college graduates.

West Indian immigrants provide another example of the importance of evaluating immigrant economic development in terms of class background and the opportunity structure of the receiving nation. West Indian incomes and occupational attainments are often cited as evidence that discrimination *per se* cannot explain the economic status of native black Americans, since so many Caribbean blacks have fared so well. In one of his many discussions of the gap between West Indian and native black American incomes, Sowell suggests the discrepancy is one more piece of evidence that "...cultural traits reaching far back in history have continuing

contemporary impact..." (1981c, p. 50). How, then, do we explain the less than sterling achievement of this same culture when functioning elsewhere—say at home in the (poverty stricken) West Indies or in Britain? Sowell acknowledges that this same culture has proven less remunerative in the West Indies (1983, p. 107), but is silent as to the reason.

Cultural explanations fail in a historical context because they consistently exclude considerations of social class. Nancy Foner (1978, pp. 229–31) reports that the Jamaicans who moved to London were of a lower socioeconomic background than their highly educated and skilled counterparts in New York City, the difference a function of respective British and U.S. immigration policies. The central point remains unaltered. West Indians in Britain and those still at home in the Caribbean share the same national culture with West Indians in the United States, but they did not share the same prior class position.

### III. Classical and Marxist Competition: Discrimination Revived

The logical problems with the cultural explanation for racial and ethnic inequality are compounded by its inadequacies in explaining the experiences of specific ethnic and racial groups. But the only reason that economic theorists have plunged into the wasteland of cultural primacy is because of the irreconcilability of market discrimination and competitive conditions. In this final section, we suggest that there are at least two conceptions of competition that can bring market discrimination back from the grave—classical and Marxist.

Classical competition (see Piero Sraffa, 1960) is rooted in the notion that there exists a tendency toward the equalization of rates of profit. Neither barriers, rigidities, nor numbers of producers in a given product market come into play so long as finance is mobile; capital must "earn" the same return in all sectors.

Suppose firms vary in productivity—with high wages in the high productivity sector, low wages in the low productivity sector and with whites in the former sector and blacks

confined to the latter. Wage differentials persist between blacks and whites of equal ability because productivity is determined by the sectoral techniques of production.

Neoclassical competition presumes labor mobility will equalize wages across sectors. If rates of return equalize, the ratio of final product prices is determined by the labor-output ratios in each sector. Austrian competition presumes entrepreneurs will move low-wage workers into the high-wage sector, initially benefiting by paying them the lower wage, but eventually eroding the wage gap. So long as rates of profits tend to equalize across sectors, these adjustments are not concerns for classical competition. Classical competition reconciles market discrimination and competition by definitionally permitting impediments or rigidities that cannot exist or persist under neoclassical or Austrian competition. Classical and neoclassical competition thus share a common inclination to define competition as a specific "state of affairs" and proceed to analyze its properties. In contrast, both Austrians and Marxists conceive of competition as a process, but differ as to the expected outcomes of the process.

Marxian competition subsumes the Classical notion, but introduces an evolutionary view (Karl Marx, 1981; John Weeks, 1981, ch. 6). Competition between capitals postulates the separation of labor from the means of production and is the means whereby the underlying laws of accumulation achieve expression. As a social relation, capital represents the integration of production and exchange in an expanding circuit, and competition between capitals arises in this integration (Weeks, p. 160). Since the market for labor power is the necessary condition for competition between capitals, the existence of capitalism implies competition (see Marx, Part 2). And, as Weeks notes, "Capitalism involves the movement of capital; competition is that movement" (p. 164).

The evolutionary nature of Marxist competition simply means that competition gives rise to monopolies, but monopoly is not the antithesis of competition. Centralization (the redistribution of existing capital) intensifies and advances competition, which is manifested in the flow of capital between branches of industry. Marx's capitalist winners con-

solidate and concentrate; they can exclude the losers and consolidate their positions for long stretches of time. Austrian competition's winning entrepreneur cannot, in contrast, *permanently* maintain barriers to preserve his position (Kirzner, pp. 131, 205). According to the Austrian view, barriers will eventually be torn down by newer contestants.

We extend Marxist competition to labor powers. Workers also can concentrate and consolidate, particularly by ethnicity or race. Via the control of training, evaluation, information, and the definition of jobs, winners in early rounds of labor market competition can insulate themselves from the most recent recruits to the wage labor force. Here lies the basis for Edna Bonacich's (1976) "split labor market," where capitalists face a labor market differentiated in terms of bargaining power and therefore price, for Oliver Cox's (1970) analysis of the racial and ethnic division of the labor force in a class-based society, and for Stanley Lieberman's queuing labor market model wherein "dominance will tend to perpetuate further dominance" (1980, p. 296).

What place remains for culture? Under classical competition with market discrimination, culture need not play any role. Austrian competition conceives of a world with entrepreneurs who heroically can render asunder all barriers. In such a world, the persistence of both market discrimination and premarket cultural differences is untenable. Marxist competition conceives of a world that tends toward monopoly. Specific ethnic and racial groups could gain control and dominance of particular occupational categories. They act to preserve their "winnings" in a fashion that is not "normal" under Austrian competition. Culture is the magnet that provides the basis for concentration of labor powers.

Extension of the Marxist conception of competition to labor powers is a relatively new theoretical endeavor, but nonetheless one with both precedent and explanatory power. Pursuit of this line of inquiry necessitates investigation of the concrete economic and historical conditions confronted by specific ethnic and racial groups. Without such an investigation, cultural analysis becomes the cornerstone for constructing what Steinberg

(p. 77-81) describes as the "New Darwinism," in which culture is passed between generations and guarantees that good things happen to the culturally efficient, and social pathology is the tragic misfortune of the cultural misfit.

## REFERENCES

- Berrol, Selma C., "Education and Economic Mobility: The Jewish Experience in New York City, 1880-1920," *American Jewish Historical Quarterly*, March 1976, 257-71.
- Bonacich, Edna, "Advanced Capitalism and Black/White Relations in the United States: A Split Labor Market Interpretation," *American Sociological Review*, February 1976, 41, 34-51.
- Chiswick, Barry R., (1983a) "An Analysis of the Earnings and Employment of Asian-American Men," *Journal of Labor Economics*, April 1983, 2, 197-214.
- \_\_\_\_\_, (1983b) "The Earnings and Human Capital of American Jews," *Journal of Human Resources*, Summer 1983, 18, 313-36.
- Cox, Oliver C., *Caste, Class, and Race*, New York: Monthly Reader, 1970.
- Darity, William A., Jr., "The Human Capital Approach to Black White Earnings Inequality: Some Unsettled Questions," *Journal of Human Resources*, Winter 1982, 17, 72-93.
- Easterlin, Richard A., "Why Isn't the Whole World Developed?," *Journal of Economic History*, March 1981, 61, 1-17.
- Foner, Nancy, *Jamaica Farewell: Jamaican Immigrants in London*, Berkeley: University of California Press, 1978.
- Freeman, Derek, *Margaret Mead and Samoa: The Making and Unmaking of an Anthropological Myth*, Cambridge: Harvard University Press, 1983.
- Kirzner, Israel H., *Competition and Entrepreneurship*, Chicago: University of Chicago Press, 1973.
- Lieberman, Stanley, *A Piece of the Pie: Black and White Immigrants Since 1880*, Berkeley: University of California Press, 1980.
- Marx, Karl, *Capital: A Critique of Political Economy*, Vol. 3, New York: Vintage Books, 1981.
- Sowell, Thomas, *American Ethnic Groups*,

- Washington: The Urban Institute, 1978.
- \_\_\_\_\_, (1981a) *Ethnic America*, New York: Basic Books, 1981.
- \_\_\_\_\_, (1981b) *Knowledge and Decisions*, New York: Basic Books, 1981.
- \_\_\_\_\_, (1981c) "Weber and Bakke and the Presuppositions of 'Affirmative Action'," in W. E. Block and M. A. Walker, eds., *Discrimination, Affirmative Action, and Equal Opportunity*, Vancouver: Fraser Institute, 1981.
- \_\_\_\_\_, *The Economics and Politics of Race*, New York: Morrow, 1983.
- \_\_\_\_\_, *Civil Rights: Rhetoric or Reality?*, New York: Morrow, 1984.
- Sraffa, Piero, *The Production of Commodities by Means of Commodities*, Cambridge: Cambridge University Press, 1960.
- Steinberg, Stephen, *The Ethnic Myth: Race, Ethnicity, and Class in America*, Boston: Beacon Press, 1981.
- Weeks, John, *Capital and Exploitation*, Princeton: Princeton University Press, 1981.
- Williams, Rhonda M., "The Methodology and Practice of Modern Labor Economics," in William A. Darity, Jr., ed., *Labor Economics: Modern Views*, Boston: Kluwer-Nijhoff, 1984.

## Women Production Workers: Low Pay and Hazardous Work

By JANIS BARRY\*

The sex segregation of the labor market is reflected in the underrepresentation of women in production jobs. In the last decade, although women have slowly entered male-intensive and better-paid production jobs, they have found greater wage inequities and more hazardous working conditions than in their traditional nonproduction jobs. Women generally have been concentrated in health-hazardous industries (such as apparel, chemical, leather, and electrical equipment), and there is significant evidence of institutional and historical forces that operate to keep them in the lower-paid production jobs within these and other industries. This has encouraged research showing that jobs women hold are differentially evaluated; this paper provides evidence of unequal rewards paid to women in hazardous jobs. Average differences in this and other job and personal characteristics between men and women are used to explain the earnings gap. A segmented labor market model is also used to derive earnings-gap explanations and to explore the importance of segment location in determining job rewards for similarly qualified workers.

### I. Conceptual Framework

During the last decade, various studies have considered the market's performance in equalizing the net advantages between jobs (Charles Brown, 1980). Neoclassical economists assert in this literature that workers are

induced into accepting jobs with disagreeable or hazardous conditions by the compensating wage differentials that employers must offer to meet competition. But due to market inadequacies that include imperfect hazard information, constrained job mobility, and underestimation of injury and disease costs, the evidence on compensating wage differentials remains inconclusive.

Segmentation theorists credit wage differentials and disparities in working conditions to the existence of not one but many labor markets, to which some workers are confined—not allocated. In effect, the neo-classical assumption that all workers are able to enforce “implicit contracts” of tradeoffs between working conditions and wages cannot hold; instead, the balance of power established between workers and their employers will resolve the question of whether working conditions degenerate, improve, or are compensated for. And bargaining power is necessarily influenced by particular historical and institutional factors. Thus, in an economy where good jobs are scarce, segmentation theorists find that otherwise qualified workers are excluded from primary-sector jobs, jobs where internal labor markets operate to promote equity and due process in the administration of work rules, high wages and uniform working conditions (David Gordon et al., 1982). This exclusion is particularly felt by women and minorities, who face strong institutional restrictions on their job choice.

While theorists disagree as to the source of labor market segmentation, many agree that primary and secondary markets are distinguished by firm characteristics and the job systems they employ. Segmentation analysis separates the primary from the secondary market on the basis of both industrial and occupational characteristics, where the meth-

<sup>†</sup>*Discussants:* Shulamit Kahn, University of California-Irvine; Luvonia J. Casperson, Louisiana State University-Shreveport.

\*Department of Social Sciences, Fordham University at Lincoln Center, New York, NY 10023.

od of categorization relies on the characteristics of jobs and not those of workers.<sup>1</sup> Production jobs in the independent primary segment are thought to provide incentives to stability in the form of high pay with some job security and rewards to general skills (this includes professional, technical, and craft jobs). Jobs in the subordinate primary segment (including blue-collar jobs in core industries) are thought to provide decent wages and advancement possibilities through internal labor market operations. Secondary jobs (including many operative and laborer jobs in peripheral industries) are thought to provide few incentives to stability because of insecure employment, low wages, few promotional possibilities, and no shelter from competitive market forces (Gordon et al.).

In previous research (1983), I found that production workers in high-risk industries and occupations receive, on average, an equalizing differential. Yet using a segmentation analysis, I found that not all personal and job characteristics are equally rewarded across segments and that compensating differentials for hazardous work are segment-specific. My earlier study also found evidence that worker-perceived hazards are most prevalent in the secondary sector, where 50 percent of all women workers are located. In fact, among all workers, women in this segment most often cited these problem-creating hazardous exposures. Thus, the question arises: do women in hazardous jobs receive compensatory wages? Further, what is the relative importance of labor-segment location in determining compensatory pay for hazardous work?

## II. Compensatory Wages for Women and Men

To specify better how the differential evaluation of jobs women hold encourages earnings inequities, I used a regression analysis to test the compensatory wage theory

using a random sample of 528 production workers. Data sources used for this investigation include the 1977 *Quality of Employment Survey (QES)* and the *Dictionary of Occupational Titles (DOT)*, from which I constructed an occupation-level hazard measure (*HAZARD*). This measure represents the mean score on six environmental conditions (cold, heat, wet, hazards, atmospheric conditions, noise) associated with the worker's occupation, as given by the *DOT*. I calculated the annual income earned on the job by each full-time worker and used the natural logarithm of this variable as the dependent variable in the earnings equation, where independent variables were worker and job characteristics.

The first regression analysis tested the alternative hypothesis that, *ceteris paribus*, there exists a positive wage differential between those workers in hazardous jobs and the wages of all other workers. The evidence shows that workers in jobs with higher mean *HAZARD* scores do indeed receive a compensatory differential. This same test was made for both men and women workers.

Table 1 shows that for men, the *HAZARD* coefficient is significant and positive. Yet the women's sample shows that women production workers earn less the more hazardous the job they hold! Comparisons between the variable means for the two samples show that the mean score on the *HAZARD* measure for women is 43 percent that of men's and, insofar as women are less likely to work in more hazardous jobs, an insignificant coefficient on this measure would seem plausible. But the unexpected finding of negative earnings premiums in hazardous jobs discredits the alternative hypothesis. In contrast to men, women are not rewarded for choosing hazardous jobs. Assuming that the women in the sample are homogeneous apart from their aversion to risk, this finding may show that women are unable to successfully exploit the same bargaining opportunities presented to men by the work environment. In Table 1, the union membership variable is significant for men but not for women, perhaps reflecting a weaker union effect on earnings within those industries where women hold more hazardous jobs.

<sup>1</sup>To avoid truncation bias, segmentation models must not base their classification of workers into segments on worker attributes such as gender, race or earnings. See Robert Buchele (1984, p. 216) for a discussion of cause vs. effects.



TABLE 1—MEANS AND REGRESSION COEFFICIENTS FOR WOMEN AND MEN<sup>a</sup>

Independent Variables <sup>b</sup>	Women		Men	
<i>HAZARD</i> <sup>c</sup>	.50	-.2104 <sup>c</sup> (.0823)	1.16	.0805 <sup>d</sup> (.0271)
<i>FRINGE BENEFITS</i>	3.56	.1093 <sup>d</sup> (.0262)	4.35	.0507 <sup>d</sup> (.0127)
<i>HAS JOB SECURITY</i>	.62	.2164 <sup>c</sup> (.0924)	.72	.0321 (.0526)
<i>UNION MEMBER</i>	.41	.1911 (.1119)	.47	.2522 <sup>d</sup> (.0476)
<i>OVERTIME</i> <sup>f</sup>	.15	.1565 (.1179)	.33	.2122 <sup>d</sup> (.0502)
<i>JOB REQUIRES SKILL</i>	.38	.0011 (.0918)	.68	.1851 <sup>d</sup> (.0517)
<i>SUPERVISOR</i>	.06	-.0359 (1.689)	.29	.1241 <sup>c</sup> (.0535)
<i>N</i> (number of observations)		72		333
<i>R</i> <sup>2</sup>		.630		.309
<i>ln</i> (annual earnings)		8.763		9.396

<sup>a</sup>Dependent variable = *ln* (1977 annual earnings). The random sample of 528 production workers is taken from the 1977 *QES*, representing workers in #401-785 and #821-824 in 1970 Census codes. Individuals with missing observations on variables entering the earnings equations are not included.

<sup>b</sup>Other independent variables used in the earnings equation are age; education; minority member; job is physically or mentally demanding; job has bad physical working conditions; repetitive work; and job tenure. These measures are described more fully in my earlier study.

<sup>c</sup>Significant at .05 level.

<sup>d</sup>Significant at .01 level.

<sup>e</sup>Mean score for six conditions.

<sup>f</sup>Minimum 10 hours per week.

In the *QES*, women production workers earned on average \$6,398 yearly and men, \$12,043. Using a decomposition procedure, we can investigate how much of the male-female earnings differential can be attributed to measured average differences in job and personal characteristics between genders. Estimations are obtained by subtracting the female mean for each independent variable from the male mean, multiplying the difference by the male regression coefficient, and expressing this project as a fraction of the differences in *ln* earnings between women and men. Of particular interest, we find that average differences between men and women on the *HAZARD* measure accounted for 8 percent of the earnings gap. Yet the standardized regression coefficients show that

the negative impact of this measure on women's earnings is almost twice the positive impact observed on men's, putting the reliability of this gap explanation in question. In general, the evidence reflects the structural constraints placed on women's pay opportunities, with job and personal characteristics cumulatively explaining only 35.2 percent of the earnings gap.

### III. Compensatory Wages within Segments

Using the same data as was used for the regression analysis for men and women, the production-worker sample was assigned to either the primary labor market (further disaggregated into the independent primary professional and technical segment, the independent primary craft segment, and the subordinate primary segment), or the secondary labor market.<sup>2</sup> A breakdown of mean *HAZARD* scores by segment shows them to be highest in the independent-craft and subordinate primary segments for both men and women. It is these workers in particular who should be earning compensating pay differentials. But separate earnings regressions by segment show that for men, secondary workers are the only group to receive a compensating differential, despite their comparatively lower *HAZARD* scores. This reflects the greater wage dispersion among men in the secondary sector, where external market forces determine earnings. Due to an insignificant sample size, separate earnings regressions for women could only be made for secondary workers; the results show a negative return on *HAZARD*, although the insignificance of the coefficient is assumed to reflect sample-size limitations. However, large earnings differences between men and women in the primary sector allows us to conjecture that women here also may realize negative returns for hazardous work because of their location in nonunionized entry-level jobs and their inability to gain the same

<sup>2</sup>The model used here (see Gordon et. al.) relies on an industry-by-occupation analysis of job characteristics, distinguishing between core and peripheral industries and those jobs which do encourage skill application.

internal labor market benefits enjoyed by men.

By comparing female-male earnings ratios within segments, I found that women in traditionally male-intensive jobs in the primary sector generally find greater pay discrimination. Women realize their greatest equity (64 percent of male earnings) in professional and technical jobs in the independent primary segment, although they represent only a scant 4 percent of all workers. Interestingly enough, women do almost as well (60 percent of male earnings) in the more female-intensive secondary sector. Yet women in craft and semiskilled jobs located in the independent primary craft and subordinate primary segments respectively earn only 43 percent and 53 percent of men's earnings.

An estimate of the cost of women's differential allocation to labor segments indicates that *ceteris paribus*, if men and women were distributed proportionally across labor segments, men would decrease their earnings by 4.9 percent and women would decrease their earnings by 7.6 percent.<sup>3</sup> An important implication of this finding is that to merely move women out of the secondary segment where they are concentrated and into the more male-intensive jobs in the primary sector would not increase their earnings unless their pay reflected an unbiased assessment of the true, relative worth of these jobs (see Robert Buchele).

Lastly, segment breakdowns of earnings differentials cumulatively explain much less of the overall gender gap because of the greater homogeneity of job and personal characteristics within segments.<sup>4</sup> But more of

the gap is explained in the secondary sector where the structure of competition largely determines the worth of women's jobs.

#### IV. Conclusions

Although compensatory wage theory stipulates that workers who assume greater risk on the job will receive additional earnings (other things being equal), my findings show that the gender composition of jobs influences the pay rates in cases of hazardous work. In general, men receive compensatory wages and women do not; specifically, women earn less the more hazardous their occupation. Women's concentration in the secondary market, where workers' bargaining power is weak, does not satisfactorily explain this, since the evidence shows that women in the more hazardous primary-sector jobs, where bargaining power is greater, generally have lower female-male earnings ratios than are found in the secondary sector. This reflects the unequal bargaining power between women and men regardless of labor-segment location. In contrast to men, the structural location of women's jobs is not as important in determining general pay rates, or (by assumption) particular compensatory pay differentials for hazardous work.

#### REFERENCES

- Barry, Janis, "Compensating Pay Differentials in Hazardous Work Situations: A Labor Market Segmentation Analysis," unpublished doctoral dissertation, New School for Social Research, 1983.
- Brown, Charles, "Equalizing Differences in the Labor Market," *Quarterly Journal of Economics*, February 1980, 2, 113-34.
- Buchele, Robert, "Sex Discrimination and Labour Market Segmentation," in F. Wilkinson, ed., *The Dynamics of Labour Market Segmentation*, New York: Academic Press, 1984.
- Gordon, David, Edwards, Richard and Reich, Michael, *Segmented Work, Divided Workers*, New York: Cambridge University Press, 1982.

<sup>3</sup>An adjusted earnings level (i.e., the antilog of the weighted average of the segment-specific ln earnings, where the weights are number of workers expected under the condition of no differential assignment to segments on the basis of gender) is compared to actual antilogged mean earnings for each gender.

<sup>4</sup>Weighting the three segment-specific gap explanations by the percentage of workers in each segment shows 22 percent of the total gap is explained for 92 percent of the sample. Earnings-gap estimates for primary professional and technical workers could not be made, but it is unlikely that this would change the total gap sum.

# Executive Compensation: Female Executives and Networking

By ROBIN L. BARTLETT AND TIMOTHY I. MILLER\*

The phrase "it's who you know, not what you know, that counts" is often heard in conversations about people who get ahead. In economic terms, this phrase translates into "it's your connections, not your human capital investments, that count." This study attempts to determine the influence of networking in addition to human capital investments by examining a sample of top female executives in the United States. We conclude that networking is as important as performance variables in helping women climb the corporate ladder.

## I. Executive Compensation: The Current Debate

To date, the primary focus of the executive compensation literature is the relative importance of firm size, as measured by sales and assets; and firm profitability, as measured by net corporate income or the rate of return on assets. The debate parallels the controversy over what is the primary objective of the modern corporation. Advocates of the neo-classical perspective believe that firms primarily maximize profits, and that more profitable firms have more productive executives, who thus earn more than executives heading up less profitable firms. Similarly, advocates of the managerial perspective argue that firms try to maximize sales given a reasonable profit margin. Thus executive compensation is directly proportional to firm sales. Although the debate as outlined by David Ciscel and Thomas Carroll (1980) continues, evidence is not compelling for

either perspective and there exists evidence that firm profits and sales are both important proxies for firm size. Although all of these studies have examined male executives, firm size and female executive compensation should be positively related.

Timothy Hogan and Lee McPheters (1980) go beyond looking at performance variables as the sole determinants of compensation, and add experience, education, and background characteristics as predictors of an executive's future performance. Typically, human capital variables such as these are not included in executive compensation models because sales and profitability variables are thought to capture an executive's contribution. As Hogan and McPheters demonstrate, however, in a world of imperfect information, personal characteristics such as experience, education, and background become important indicators of future productivity as suggested by the screening hypothesis or favorable training costs as implied by the job-competition hypothesis. A study by Naresh Agarwal (1981) compares the relative importance of human capital variables and organizational factors such as the layers of management and the firm's ability to pay. Agarwal concludes that human capital factors may be important for initial entrance, but organizational factors dominate personal characteristics thereafter.

Additional reasons for including human capital variables are: 1) ascribing a firm's performance totally to the performance of the chief executive officer is problematic; and 2) none of the women in our sample is the chief executive officer of a Fortune 500 firm. Thus, using the human capital theory (Gary Becker, 1964), an individual's compensation is directly related to her productivity; that is, in a competitive setting, an individual's wage is equal to the marginal net revenue product associated with her employment. Moreover, an individual's remuneration depends upon other factors in addition to her initial pro-

\*Professor of Economics and Associate Professor of Economics, Denison University, Granville, OH 43023. We thank Jacqueline Davis and Korn/Ferry International for making their sample of executive women available, and the Wellesley College Faculty Awards Committee for providing the funds necessary to secure it. We also thank Daniel Fletcher, Judy Thompson, and Sara Jean Wilhelm for editorial comments.

ductive capacities. Compensation grows over time as an individual accumulates human capital through additional formal education or informal on-the-job training. An executive with more experience, more and/or better education, and more general and specific training, *ceteris paribus*, should perform better on the job and earn a higher salary.

## II. The Female Executive

A great deal of attention is being focused on the new managerial women. The reason for such attention is that women are moving very rapidly into the managerial ranks. In 1961, 5.1 percent of the female labor force was in the managerial and administrative occupation. By 1981, the percent of the female labor force in that occupation had grown to 7.4 percent. The percent of the male labor force hovered around 13.5 percent over the same time period. The influx of women into the managerial ranks increased gender concentration (the percent of an occupation female) from 15 percent in 1961 to 27 percent in 1981 (U.S. Department of Labor, 1983). But, the income statistics do not indicate that managerial women are getting ahead. Female managers earned 58 percent of what male managers earned in the early 1960's, and twenty years later the ratio remains virtually unchanged (U.S. Department of Commerce, 1963; 1983). Firm profitability and size may account for this difference, human capital variables may account for this difference, female managers may be differentially motivated than their male counterparts, or female managers may have relatively limited access to labor market information. However, Linda Brown (1979) concludes in her survey article that there is no evidence of consistent motivational differences between male and female managers. Successful managers tended to be ambitious, willing to take risks, and the oldest child of a professional father. In a recent case study of two major retail firms, one with 19 percent and the other with only 6 percent of its managers female, Anne Harlan and Carol Weiss (1981) examined the unique background characteristics of these women and, the organizational gender biases of their

firms. While the managerial women in this study did not exhibit any motivational differentials as compared to their male counterparts, they did believe that being a woman was the biggest hindrance to their career. Both men and women in this study felt that women were given different training, development opportunities, promotional criteria, and supervision than men were. The interaction of individual characteristics and organizational gender biases resulted in limited advancement opportunities for women. If motivational and background characteristics do not explain the persistent wage gap, then information differentials need to be explored.

## III. Networks in a World of Imperfect Information

Without perfect information in the labor markets; that is, without buyers and sellers of executive talent having complete information, both groups need to engage in search activity. Albert Rees (1966) discusses the role of formal and informal information networks in labor markets. Since it is difficult for an executive to be knowledgeable about all the job possibilities in the market, she lists her name with an executive recruiting firm, and when a job possibility arises, pursues it. "Head-hunting" firms are an example of a formal channel of information. Friends and family are an informal source of labor market information. Although not as formal as an executive recruiting firm, nor as informal as family, clubs are a valuable source of labor market information. Social or service clubs tend to draw their members from similar socioeconomic backgrounds. Belonging to a large network of similarly situated professionals offers a greater opportunity for hearing about jobs (the extensive margin of job search) and a greater opportunity for finding out inside information about a job offer (the intensive margin of job search). Mary Corcoran, Linda Datcher, and Greg Duncan (1980) found that over 40 percent of all managerial jobs are discovered by word of mouth. Information is efficiently pooled and cheaply distributed through club affiliations. An executive who belongs to clubs would be expected to have better infor-

mation at both the extensive and intensive margins of job search. As George Stigler notes: "The information a man possesses on the labor market is capital: it was produced at the cost of search and it yields a higher wage rate than on average would be received in its absence" (1962, p. 103). In our study on the determinants of executive compensation with James Grant (1984), we found that male executives who belonged to clubs earned more than nonaffiliated executives. Female executives who belong to clubs should also earn more than female executives who do not.

#### IV. The Model

The model as developed (limited by the available data), is as follows:

$$(1) \quad Y = b_0 + b_1S + \sum (b_iH) + \sum (b_iB) \\ + \sum (b_iO) + \sum (b_iN) + U.$$

Equation (1) says that an executive's compensation  $Y$ , which equals salary plus bonuses, depends upon the size of her corporate employer  $S$ , her human capital investments  $H$ , her unique background characteristics  $B$ , her (perceived) organizational environment  $O$ , and her access to information networks  $N$ . The  $b_i$ s are the coefficients to be estimated and  $U$  is the disturbance term. The actual variables used in each category are listed in Table 1. Differences in these variables will lead to differences in individual compensation.

#### V. The Sample

In 1982, a questionnaire was designed by UCLA Graduate School of Management and Korn/Ferry International, an executive recruiting firm, to obtain a profile of the successful female executive. The questionnaire was sent to over 600 female executives from Fortune's top 1000 firms, and more than 50 percent were returned. Our sample consists of the 132 female executives who responded to those questions that were essential to our study. Some of the more interesting char-

acteristics of these women are: their average cash compensation is \$86,881; their average age is 41.9 years; overall, 26.5 percent graduated from an "Ivy League" college, 43.1 percent held some leadership position in college, and 13.8 percent participated in sports while in college. On average, these women worked 52.5 hours per week and 26.2 percent took a leave of absence during their career. Less than half, 48.2 percent, are currently married; 53.4 percent had fathers who were professionals; and 56.0 percent were the oldest child. When asked to rate their level of ambition on a scale of 1 to 5 (1 being least), the average value given was 4.207; 65.2 percent viewed themselves as "risk takers," and 47.0 percent are Republicans. Of particular interest to this study, 56.3 percent are members of private clubs and 20.9 percent serve on corporate boards. In terms of progress up the corporate ladder, 41.6 percent believe that "it's who you know, not what you know" that matters.

#### VI. The Regression Results

The model was estimated using both the linear and semilog specifications. In estimating our model, several persistent cases of multicollinearity occurred. Age, age-squared, and job tenure were gathered as measures of human capital, but because these women executives were all very similar, it was not possible to include more than one of these variables in the model. Age was chosen over age-squared, since the sample consisted of relatively young executives who were probably not at the peak of their careers. Age and job tenure were then included separately. The variable job tenure had a  $t$ -score less than .06, while age had a  $t$ -score greater than one. Since both could not be used, age was chosen. A similar problem arose in attempting to use "college degree," "Ivy League" college graduate, "Masters of Business Administration," and "held a college leadership position" as separate variables. The variables "held a college leadership position" and "Ivy League" college graduate were selected since the former measured an important antecedent background characteristic and the latter

TABLE 1—DETERMINANTS OF FEMALE EXECUTIVE COMPENSATION

Dependent Variable	Cash Compensation	Ln Cash Compensation
The Size of the Corporation, <i>S</i>		
Size of employer's corporation	1.976 (1.395) <sup>a</sup>	.0324 (2.726) <sup>b</sup>
Human Capital Investments, <i>H</i>		
Age	.489 (1.007)	.0044 (1.083)
Ivy League Graduate	17.321 (1.946) <sup>b</sup>	.1397 (1.874) <sup>b</sup>
Number of Hours Worked	1.322 (2.417) <sup>b</sup>	.0101 (2.191) <sup>b</sup>
Unique Background Characteristics, <i>B</i>		
Married	8.100 (1.028)	.1075 (1.629) <sup>a</sup>
Father was a Professional	-2.591 (-.325)	-.0086 (-.128)
Oldest Child	-2.245 (-.290)	.0109 (.169)
Views Self as Ambitious	6.264 (1.475) <sup>a</sup>	.0615 (1.728) <sup>b</sup>
Takes Risks	11.322 (1.332) <sup>a</sup>	.0960 (1.347) <sup>a</sup>
Held College Leadership Position	4.703 (.573)	.0831 (1.208)
Active in College Sports	16.107 (1.425) <sup>a</sup>	.1086 (1.147)
Not a Republican	6.693 (.834)	.1205 (1.792) <sup>b</sup>
Organizational Environment, <i>O</i>		
Believes Barriers have Fallen	-.542 (-.147)	-.0003 (-.009)
Perceives Discrimination	-8.669 (-2.363) <sup>b</sup>	-.1017 (-3.309) <sup>b</sup>
Networking, <i>N</i>		
"Who you know"	-14.623 (-1.797) <sup>b</sup>	-.1053 (-1.545) <sup>a</sup>
Member of Corporate Board	38.987 (3.962) <sup>b</sup>	.3120 (3.784) <sup>b</sup>
Private Club Member	21.923 (2.771) <sup>b</sup>	.2515 (3.796) <sup>b</sup>
Intercept	-41.637 (-.925)	3.1650 (8.398) <sup>b</sup>
<i>R</i> <sup>2</sup>	.349	.407
<i>F</i>	3.596 <sup>b</sup>	4.593 <sup>b</sup>
<i>N</i> = 132		

<sup>a</sup>Indicates significance at the 10 percent level using a one-tail test.

<sup>b</sup>Indicates significance at the 5 percent level using a one-tail test.

measured educational credentials. Finally, the variables "took a leave of absence" and "number of children" were also considered as measures of unique background characteristics. These variables were statistically insignificant (*t*-scores were -.068 and .069,

respectively), and their presence in the model added substantially to the multicollinearity problem. The results are presented in Table 1.

Because cross-sectional data were used, heteroscedasticity was suspected as in previous studies. Both a Glejser test and a Goldfeld-Quandt test as outlined in John Johnson (1984, pp. 300-02) were performed. No evidence of heteroscedasticity was detected at the 10 percent level with either test.

## VII. Interpretation of Results

These results are a mixture of the anticipated and unexpected. Executive compensation is positively influenced by both the size of the corporation and the amount of effort put forth by the individual (as measured by the number of hours she works per week). The quality and/or quantity of one's formal education (as measured by "Ivy League") seems to have a positive effect upon compensation although experience (as measured by age) was not significant.

Contrary to our expectations, marriage was not an inhibiting factor for female corporate success. Two of the more disappointing results were the statistical insignificance of being an only child and having a professional father. If one considers that 56 percent of these executive women were oldest children and that over 53 percent had professional fathers, it seems improbable that these traits do not matter. Perhaps these factors would better serve to differentiate between women who become executives and those who do not. Neither being married nor being a non-Republican had the anticipated negative impact. Having been active in college sports had a positive impact on compensation, suggesting that this variable may be an indicator of the young woman's willingness to take risks and/or her learning to be a "team player."

These results strongly support Harlan and Weiss in their contention that organizational gender biases play a major role in limiting the advancement and compensation of female executives. Working in an environment in which discrimination is readily perceived has

a very strong negative impact upon a woman's compensation, no matter what her attributes or efforts. While it was noted earlier that women are entering management in increasing numbers, these results provide evidence that their progress up the corporate ladder is being impeded by other barriers.

The major premise of this paper is that "who you know" may be just as important as "what you know." This contention seems to be strongly supported by the results. Working environments in which knowing the right person is important for career advancement has a strong negative impact upon a female executive's compensation. In contrast to this, women who "plugged in" to networks, as measured either by memberships in private clubs or on corporate boards, profited substantially by these additional contacts and information. Thus, managerial women need more than larger corporations and greater human capital investments to be successful. In order to make the most of their opportunities and human capital investments, women need to be plugged in to networks that provide requisite labor market information.

#### REFERENCES

- Agarwal, Naresh C., "Determinants of Executive Compensation," *Industrial Relations*, Winter 1981, 20, 36-45.
- Bartlett, Robin L., Grant, James H. and Miller, Timothy I., "Executive Compensation: Systematic Risk and Differentials by Executive Type," Working Paper No. 82, Department of Economics, Wellesley College, 1984.
- Becker, Gary S., *Human Capital*, New York: National Bureau of Economic Research, 1964.
- Brown, Linda K., "Women and Business Management," *Signs*, Winter 1979, 5, 266-89.
- Ciscel, David H. and Carroll, Thomas M., "The Determinants of Executive Salaries: An Econometric Survey," *Review of Economics and Statistics*, February 1980, 62, 7-13.
- Corcoran, Mary, Datcher, Linda and Duncan, Greg J., "Most Workers Find Jobs through Word of Mouth," *Monthly Labor Review*, August 1980, 103, 33-35.
- Harlan, Anne and Weiss, Carol, "Moving Up: Women in Managerial Careers," Working Paper No. 86, Wellesley College Center for Research on Women, 1981.
- Hogan, Timothy D. and McPheters, Lee R., "Executive Compensation: Performance Versus Personal Characteristics," *Southern Economic Journal*, April 1980, 46, 1060-68.
- Johnson, John, *Econometric Methods*, 3rd ed., New York: McGraw-Hill, 1984.
- Rees, Albert, "Information Networks in Labor Markets," *American Economic Review Proceedings*, May 1966, 56, 559-66.
- Stigler, George J., "Information in the Labor Market," *Journal of Political Economy*, October 1962, 70, 94-105.
- U.S. Department of Commerce, Bureau of Census, *Current Population Reports: Consumer Income*, Series P-60, Washington: USGPO, various issues.
- U.S. Department of Labor, *Employment and Training Report of the President*, Washington: USGPO, 1983.

# Longitudinal Changes in Salary at a Large Public University: What Response to Equal Pay Legislation?

By SHARON BERNSTEIN MEGDAL AND MICHAEL R. RANSOM\*

Legislation in the early 1970's greatly strengthened the legal status of women faculty at colleges and universities. The Equal Employment Opportunity Act of 1972 extended Title VII of the Civil Rights Act protection against sex discrimination in employment to academic personnel. In the same year, the Equal Pay Act of 1963 was expanded to cover university and college faculty, and orders from the Executive Branch required affirmative action plans of all federal contractors. The purpose of this paper is to determine the extent to which the salary structure at a large public university has changed since 1972. In particular, we use longitudinal data to examine how the University of Arizona responded to this legal environment with regard to employment and pay of female faculty members.

Many have gauged sex differentials in faculty compensation and offered reasons for their existence. The general empirical findings of these studies can be summarized as follows. After controlling for numerous factors, women tend to earn lower salaries than do men, and the difference is usually statistically significant. The differential tends to widen with length of service, while controlling for rank lowers its size, although rank assignment itself may reflect sexual discrimination in hiring and promotion (see Emily Hoffman, 1976; Burton Malkiel and Judith Malkiel, 1973). With few exceptions, most have used cross-sectional data for a single year. Changes in compensation practices in response to equal pay legislation have not been assessed. From a longitudinal point of view, the "wage gap" between com-

parable male and female faculty members at a given time can be thought of as arising from two different sources: differences in pay at the time of hire; and differences in rates of growth in salaries. If there is an institutional response to the changing legal environment, it must appear as a change in either or both of these factors. In what follows, we describe our data set and summarize our findings. As is suggested by our title, we find that significant unexplained salary differentials have persisted.

## I. Data

Our data set contains information on salaries and personal and employment characteristics of about 1900 individuals who were employed by the University of Arizona during 1972, 1977, and/or 1982. We restrict our analysis to career-type teaching faculty at the main campus of the university; administrators, professional and research faculty, and medical school faculty are excluded from the data set. We were able to identify salary, rank, tenure status, and departmental and collegial affiliation from the annual budgets of the university. From other published university records, we determined the date and type of highest degree and the date of hire. Sex was determined by name. In the case of unusual or ambiguous names, the sex was verified through personal communication. Some individuals are not included in our sample because of incomplete information.

From the raw data, we construct for each year the following variables: *SEX*, *PHD*, *SERV*, and *EXPR*. The variable *PHD* is a dummy equal to one if the individual has a doctoral degree, such as a Ph.D. or Ed.D. The variable *SERV* is defined as the years of service at the University, and *EXPR* is defined as the number of years since the individual received his or her highest degree.

\*Department of Economics, University of Arizona, Tucson, AZ 85721. We thank Cordelia Reimers and Michael Rieber for helpful comments.



TABLE 1—UNIVERSITY OF ARIZONA MEAN SALARY AND EMPLOYMENT DATA FOR ALL COLLEGES AND THE LIBERAL ARTS COLLEGE BY SEX

	<i>SAL</i>	<i>LNSAL</i>	<i>PHD</i>	<i>SERV</i>	<i>EXPR</i>
<u>1972</u>					
All Colleges					
Overall	16304	9.68	.75	9.3	13.1
Male	16702	9.70	.80	9.2	13.0
Female	13675	9.51	.42	10.4	14.0
Liberal Arts					
Overall	16169	9.67	.91	9.2	12.9
Male	16441	9.69	.94	8.8	12.6
Female	13328	9.48	.68	13.2	15.8
<u>1977</u>					
All Colleges					
Overall	22060	9.97	.82	10.2	13.9
Male	22672	10.00	.85	10.5	14.4
Female	18385	9.80	.64	8.6	11.0
Liberal Arts					
Overall	22038	9.97	.92	10.1	14.2
Male	22527	9.99	.94	10.2	14.6
Female	18005	9.78	.79	9.1	11.4
<u>1982</u>					
All Colleges					
Overall	34088	10.41	.84	12.8	17.1
Male	35158	10.44	.86	13.2	17.6
Female	27406	10.20	.69	10.6	14.0
Liberal Arts					
Overall	34241	10.41	.93	12.8	17.5
Male	35100	10.43	.94	13.2	18.0
Female	27374	10.19	.86	9.6	13.2

*SEX* is a dummy variable equal to 1 for females; *SAL* is the individual's salary, adjusted to a 10-month academic year, and *LNSAL* is the natural logarithm of salary.

Table 1 presents summary statistics for the basic variables of our analysis for the full university and for the Liberal Arts College. We chose to focus on a single college within the university because colleges have a lot of autonomy in employment decisions. The Liberal Arts College was selected for its relatively large number of women. Corresponding sample sizes *N* (total) and *NF* (females) are reported in Table 2. There are rather striking differences between males and females. Male salaries average 22 to 28 percent more than female salaries, with the ratio increasing over time. Over the same period, average *EXPR* and *SERV* of females have declined substantially relative to male averages. This is due primarily to the much higher rate of turnover among females. For exam-

TABLE 2—UNEXPLAINED SEX DIFFERENCES IN SALARIES: FULL SAMPLES AND RECENT HIRES<sup>a</sup>

	<i>N</i>	<i>NF</i>	Levels	Logs
<u>All Colleges</u>				
Full Sample				
1972	1027	135	-1549 (5.72)	-.105 (6.51)
1977	1067	152	-1191 (3.48)	-.063 (4.31)
1982	1137	157	-2853 (5.19)	-.095 (6.17)
Recent Hires				
1972	226	42	-1353 (3.01)	-.088 (3.06)
1977	208	43	-1237 (1.97)	-.059 (2.10)
1982	170	44	-2245 (2.18)	-.086 (2.60)
<u>Liberal Arts</u>				
Full Sample				
1972	389	34	-2321 (4.83)	-.151 (5.45)
1977	444	48	-1647 (3.20)	-.084 (3.95)
1982	459	51	-3595 (3.75)	-.116 (4.46)
Recent Hires				
1972	83	9	-1724 (1.96)	-.098 (1.82)
1977	72	9	-2167 (2.34)	-.106 (2.53)
1982	55	12	-2566 (0.97)	-.099 (1.24)

<sup>a</sup>Absolute values of *t*-statistics are shown in parentheses.

ple, of the females present in 1972, 47 percent left before 1977 and 66 percent before 1982. The corresponding rates for men were 29 and 41 percent. Another consequence of the high turnover is that while more than 20 percent of the "recent hires" are female, the fraction of females in the full sample in each period is a steady 13 percent.

## II. Empirical Findings

Tables 2 and 3 present the male-female differentials in salary and growth of salary which remain after controlling for *PHD*, linear and quadratic measures of *SERV* and *EXPR*, and the interactive variables *PHD* × *EXPR*, *PHD* × *SERV*, and *EXPR* × *SERV*. We ran each regression for all the colleges combined, in which case we control for col-

TABLE 3—UNEXPLAINED SEX DIFFERENCES  
IN SALARY GROWTH<sup>a</sup>

	<i>N</i>	<i>NF</i>	Levels	Logs
<b>All Colleges</b>				
1972-77	660	71	-41 (0.85)	.019 (2.25)
1972-82	557	47	-583 (0.87)	.037 (2.23)
1977-82	768	84	-629 (1.61)	.012 (1.20)
1977-82 <sup>b</sup>	541	45	-358 (0.70)	.017 (1.40)
1977-82 <sup>c</sup>	227	39	-910 (1.42)	.004 (0.25)
<b>Liberal Arts</b>				
1972-77	274	22	-460 (1.40)	.012 (1.05)
1972-82	237	14	-949 (0.78)	.049 (1.71)
1977-82	338	29	-366 (0.56)	.022 (1.43)
1977-82 <sup>b</sup>	237	14	-66 (0.08)	.043 (2.25)
1977-82 <sup>c</sup>	101	15	-990 (0.84)	-.011 (0.35)

<sup>a</sup>Absolute values of *t*-statistics are in parentheses.<sup>b</sup>1972 cohort.<sup>c</sup>1977 cohort.

legal affiliation, and for the Liberal Arts College, in which case we control for departmental affiliation.

Table 2 lists the unexplained salary differentials (the coefficients for *SEX*) by year for all the faculty employed in that year (Full Sample) and for those hired in the preceding three years (Recent Hires). Note that the recent hires group includes all those recently hired rather than only those hired at entry level positions. We report the differentials for the two alternative measures of the dependent variable, *SAL* and *LNSAL*. Thus, we have estimates of the average salary difference in both dollar and percentage terms. Though we see a dip in the size of the differential in 1977, the differentials are negative and significant at a 5 percent significance level in most cases. The regressions for the full sample explain between 55 and 70 percent of the variation in the dependent variable, while the recent hires regressions explain a somewhat larger percentage.

Some additional analyses were done to determine the effect of controlling for rank.

The adjusted *R*<sup>2</sup>s increase by 10 or more points, but the full sample sex differentials, while smaller in absolute value (by a factor of .2 to .6), are still significant. The significance of the recent hires differential varies by year and subsample. Thus, even if we assume no sexual bias in rank assignment, controlling for rank does not eliminate the sex differential in compensation. Moreover, it does not eliminate the rise in the salary gap between 1977 and 1982. We also ran the regressions with department heads included, with similar results. Department heads receive significantly higher salaries. Since none of the 64 department heads were female in 1972 and 1977, and only 4 of the 70 department heads were female in 1982, these findings indicate that additional salary discrimination may occur through bias in assignment of administrative positions. We also tested for equality of male and female full sample structures, where the unrestricted regression allowed for different structures. The restricted regression constrained all coefficients to be equal except the intercept; that is, the *SEX* variable was included. We accepted the null hypothesis of homogeneous Liberal Arts structures at a 5 percent significance level, but we rejected the null hypothesis when all the colleges were combined, which suggests that more accurate estimates of the sex differential are obtained when we control for departmental affiliation. Finally, regressions for colleges other than Liberal Arts yield sex differential estimates which are sometimes insignificant or positive, a finding that explains why the Liberal Arts College sex differentials are larger than the full university differentials.

Table 3 reports the unexplained sex differences in salary growth over the periods listed in the first column of the table. The sample sizes (*N*) reflect the number of faculty employed in both years of the period. In addition, we partition the 1977-82 sample into two groups: those employed in 1972, 1977, and 1982 (1972 cohort) and those not employed in 1972 (1977 cohort). The explanatory variables are assigned values according to the starting year of the period. In dollars, female faculty have received smaller salary increases, on average, but the dif-

ference is not statistically significant. Moreover, female salaries have grown faster, on average, than male salaries, and the difference is sometimes significant. The results for the 1977–82 period suggest, however, that only the cohort of female faculty present in all three years benefited from the more-rapid salary growth. Nevertheless, the growth was not sufficient to eliminate the sex differential. For the cohort of faculty present since 1972, the unexplained sex differential in 1982 is –6.6 percent and is statistically significant.

### III. Conclusions

We have found that, after controlling for a limited set of individual characteristics, female faculty at the University of Arizona are paid less than male faculty. These differences persist over time. Yet, we find that for those women who were present at the university over the entire period studied, the gap has not worsened. In fact, female salaries grew at a higher rate than male salaries. Thus, it appears that most of the differential in salaries is due to the treatment of females at the time of hire. (Our recent-hires regressions also reflect this.) This is a curious result, since it is precisely at the time of hire, when individuals are most mobile, that there is least opportunity to discriminate. The university must pay competitive market salaries to new hires—any differential at this point must be due to differences in unmeasured qualifications (productivity) and/or discrimination in the market-at-large. One possible explanation for lower starting salaries could be that due to implementation of affirmative action plans nationwide, highly qualified women are hired in preference to highly qualified men by the most prestigious universities and by industry. This leaves a

pool of applicants for the remaining institutions in which women are relatively less qualified than men. Alternatively, the salary and turnover differentials we observe could reflect the university's lack of commitment to sex equity in faculty recruitment, promotion, and retention.

Both the market sorting and lack-of-commitment explanations support the rather unsettling conclusion that the salary differences represent quality differences. Their implications regarding compliance with nondiscrimination legislation, however, are quite different. The explanation that the differentials reflect discrimination in the market-at-large still remains. This explanation is certainly consistent with the findings of others. It is also consistent with our finding that the salary differential dipped in 1977, only to rise again in 1982. In response to 1972 legislation threatening to deny federal program funding to institutions guilty of sex discrimination, universities across the country adopted affirmative action programs and sought to hire more women faculty. By the late 1970's, however, the clamor appeared to have subsided. Although our results suggest little response to the legislation of the early 1970's by 1982, recent actions at the University of Arizona are encouraging. The administration voluntarily awarded women faculty \$300,000 in salary adjustments for 1985.

### REFERENCES

- Hoffman, Emily, "Faculty Salaries: Is There Discrimination by Sex, Race and Discipline? Additional Evidence," *American Economic Review*, March 1976, 66, 196–98.
- Malkiel, Burton G. and Malkiel, Judith A., "Male-Female Pay Differentials in Professional Employment," *American Economic Review*, September 1973, 63, 693–705.

# Sex Role Socialization and Labor Market Outcomes

By MARY E. CORCORAN AND PAUL N. COURANT\*

It is well known that women earn less than men. The causes of the wage gap, however, are not well understood. Each of the three major economic explanations of the wage gap—human capital, pure discrimination, and “crowding”—suffers from serious shortcomings. The human capital explanation is simply not supported by the data: sex differences in experience, training, and work history can account for only about one-third of the wage differences between men and women. (See Corcoran and G. J. Duncan, 1984, and references cited therein.) Pure discrimination (especially to the tune of about 25 percent of male wages) is just not sustainable in a competitive market (Kenneth Arrow, 1972a, b). And although women are very much crowded into a relatively few occupations (Barbara Bergmann, 1974), this fact does not carry with it an explanation of how and why it comes to be true.

Socialization and discrimination are the two explanations of the unexplained portion of the wage gap that are most often identified (see, for example, D. J. Treiman and H. I. Hartmann, 1981), but the discussion usually stops there. Our purpose in this paper is to begin to look explicitly at how socialization might work in a model of the labor market, how socialization and discrimination might interact, and how one might test empirically for labor market effects of socialization.

## I. Socialization

In an extensive review of the literature, J. P. Eccles and L. W. Hoffman (forthcoming) suggest that sex differences in socialization might affect occupational behavior in at least four ways. First, socialization may

lead women to be more fearful or more anxious, or less confident than men are (the “fear of success” syndrome). Second, sex role socialization may directly affect workers’ skills and personality traits. Some researchers argue, for instance, that girls are encouraged to be more dependent, more person-oriented, and less able mathematically than are boys. Third, children may internalize traditional notions of sex roles, accept these cultural sex stereotypes as fact, and eventually choose occupations that conform to these stereotypes. Fourth, sex role socialization may affect the values men and women attach to different activities so that workers of both sexes tend to value “sex-appropriate” activities. Thus, women may value person-oriented tasks more than men do, even if there were no sex differences in ability to perform such tasks.

The first two sets of phenomena are really human capital arguments. In both cases, women differ from men in ways that might reduce these women’s potential value in the labor market. Such sex differences in human capital may or may not have been caused by discrimination. The third and fourth findings suggest that equally qualified men and women may evaluate the same job characteristics quite differently when choosing jobs. They are thus “taste” explanations when considered from the perspective of the labor market.

## II. Modelling Strategy

If we are to be able to distinguish among various kinds of discrimination and socialization as explanations of male-female pay differentials, we need a model that allows for a variety of possibilities and interactions among them. Given our introductory discussion, such a model must include the following: 1) labor market discrimination against women; 2) socialization of women that tends to lead women to value occupations and jobs that are viewed as “traditionally female” in

\*Institute of Public Policy Studies, 1516 Rackham Building, University of Michigan, Ann Arbor, MI 48109.

character; 3) phenomena (including socialization of employers and customer and co-worker prejudice) that lead employers to value a workplace in which roles are performed by employees of traditionally appropriate gender; and 4) interactions among 1), 2), and 3).

We place one more requirement on a model that purports to shed some light on the issue of how it is that women come to be paid so much less than men. The model must account for the fact that there is a great deal of heterogeneity in the tastes and socialization of workers of both sexes. Thus a convincing account of what is going on cannot depend on the notion that all women, or all employers, have uniform views on the sex appropriateness of different occupations. Further, to the extent that there are employers who do not discriminate in any way on the basis of gender (and surely there are), any elements of a model that depend on employer discrimination must explain why the competitive process does not cause such behavior to disappear.

Throughout this discussion, we assume that the labor market is competitive in the sense that no one possesses monopoly or monopsony power, and that markets for all types of labor are "thick." We assume that individual workers (or potential workers) have utility functions that are defined on income ( $Y$ ), leisure ( $L$ ), and attributes of their job (a vector  $Z$ ). Further, we define a function  $F(Z)$  that maps the vector  $Z$  into a vector of indexes of traditional sex appropriateness for each element of  $Z$ . The more consistent with traditionally female roles, the higher the value of any element of  $F$ . Thus, for any worker  $k$  employed in job  $j$ , the utility function is

$$(1) \quad U_k = U_k(Y, L, Z^j, F^j(Z^j)).$$

Worker  $k$ , of course, will be in that job  $j$  which maximizes (1) subject to the set of jobs and associated wages available on the market. Worker  $k$  may not care about the vector  $F^j$ , but then again he or she may care about it. On average, the partial derivatives  $\partial U / \partial f_i$  (where  $f_i$  is an element of  $F$ ) are positive for women and negative for men. Note that attributes of the job may include such elements

as the sex ratio of current employees, the extent of perceived co-worker prejudice, and the degree to which such prejudice would directly affect the worker.

Sex appropriateness may also matter on the demand side of the labor market. Thus the wage offered by employer  $m$  to potential employee  $k$  will depend on  $k$ 's human capital characteristics ( $Z^k$ ), the attributes of the job, and the vector  $F^j$ . Thus, for employer  $m$  considering employee  $k$  in job  $j$ , the offered wage is

$$(2) \quad w_m^{kj} = w_m(Z^j, F^j(Z^j), Z^k).$$

Again, employer  $m$  may care about elements of  $F^j$  or may not, and may also be affected (following Arrow) by co-worker and customer prejudice, even if the employer himself (herself) has no prejudice.

In a perfectly competitive market, direct employer discrimination will not be sustainable, although if it is difficult for employers to identify prejudiced workers, and if such workers are less productive when they work with the objects of their prejudice, it is still possible that employers would systematically pay one sex less than the other in certain jobs, even if the human capital characteristics of employees were identical.<sup>1</sup> Any differences in wages as a function of gender other than those attributed to co-worker discrimination would be due to differences in employee tastes. If (following the psychological literature cited above) jobs with attributes that are valued by women more than by men are relatively scarce, then it is easy to tell a story consistent with Bergmann's crowding hypothesis that would explain why jobs that are largely filled by women pay less (for a given vector of measured human capital attributes) than jobs that are mostly filled by men. All that is occurring is market equilibrium with a compensating wage differential. Moreover, to the extent that co-worker discrimination affects wages by tending to

<sup>1</sup>The proof of this and other assertions that are not proved in the text due to the limitations placed on the length of this paper are available from the authors, as are the data sources.

reduce the supply of women to jobs where co-workers would hassle women, but where output would be unaffected by sex integration, crowding of women into occupations where they were not faced by such behavior could also be interpreted as yielding compensating differentials—the overt discrimination is not on the part of the employer.

Formal models of the labor market that use (1) and (2) as building blocks can be used to analyze a large number of ways in which prejudice, socialization, and discrimination can interact to produce wage differentials by gender. One interesting version of such a model involves employee search over jobs, and under conditions of costly search leads to the possibility that perceptions of discrimination (whether real or not) can interact strongly with tastes on the part of employees regarding sex appropriateness in ways that can lead to large equilibrium wage differentials. Here equilibrium means that it will not be rational for a woman, earning typical women's wages in a traditionally female line of work, to search for work in the male labor market, where wages are higher, but so is the probability of being hassled on the job, being uncomfortable with performing gender-inappropriate activities, or simply being denied serious consideration for the job.<sup>2</sup>

The value of models constructed along these lines lies not in the fact that they yield any startling new predictions, but in the fact that they provide a notation in which it becomes abundantly clear that at the level of "standard" human capital earnings functions, explanations of male-female wage differences that are based on differences in employee tastes for job attributes are indistinguishable from explanations that are based on employer (or co-worker) discrimination. It is fairly easy to write down models in which

lots of potentially important factors interact, but if we are to find out which of these factors are important in the labor market, we need empirically testable implications of (say) socialization-based stories that are different from discrimination-based stories.

### III. An Empirical Research Agenda

The ideal way to disentangle the possibilities implicit in the above discussion would involve following a panel of children over time and examining how their family environments and their school environments affect their sex-role attitudes and aspirations; how families, schools, attitudes, and aspirations influenced decisions about investment in education and training; and how families, schools, attitudes, aspirations, and human capital affected job choice and wages. At key decision points (choice of college major, first job, etc.), we would need to ask detailed questions about the factors which influenced those decisions—particularly about paths not taken. We know no data that would allow us to take this approach.

We can, however, derive some estimate of how powerful socialization effects might be by looking at specific indicators of socialization and seeing how these fare in the labor market. For instance, a key feature of a socialization explanation model is that sex-role patterns learned in childhood socialization will affect adult economic behavior. Psychological studies have identified the following as family factors which promote sex-role differentiation among children: being raised in a female-headed household, being raised in family with children of one sex, having nontraditional parents (see Eccles and Hoffman; M. M. Marini and M. C. Brinton, forthcoming, for summaries of this research). We are currently using data on 800 young women aged 25–30 in 1981 to test directly for a link between such family factors and young women's market outcomes. These data are taken from the Panel Study of Income Dynamics (PSID), a 14-year (1968–81) study of a nationally representative sample of American families. These young women were children aged 12–17 in their parents' homes in 1968 and had established their own homes

<sup>2</sup>Note that not all women have to have the same perceptions or tastes regarding gender appropriateness. All that is required to get average differences in pay is that women on average value traditional female roles more than men do, and dislike discrimination and hassle from co-workers.

by 1981. We are also using a sample of brother-sister pairs from the PSID to isolate families that maximize similarity (dissimilarity) between brothers' and sisters' labor market outcomes. Using these outliers, we will try to identify family factors that account for this similarity (dissimilarity).

Even if these PSID analyses establish a link between family factors and young women's labor market outcomes, they will not provide much information about the processes by which family factors affect economic outcomes. The PSID provides virtually no data on sex-role attitudes—and a key prediction of socialization models is that women who value traditional roles will be more likely to choose "female" jobs. We will use the General Social Survey (GSS), a nationally representative cross-sectional data set with excellent measures of sex role attitudes, to test whether such attitudes influence women's labor supply, occupations, and wages independently of education and training.

A third source of data, the Thornton Longitudinal Study of Families, has much richer detail on both family background and attitudes for a sample of 906 mother-child pairs from Detroit over the period 1962–80. This study has six waves of data and provides extensive information about mothers' sex-role attitudes (measured 3 times between 1962 and 1980), mothers' work histories, the household division of housework, children's sex-role attitudes in 1980, children's test scores and schooling, and children's aspirations and early labor market outcomes. We plan to use these data for a more complete explanation of how family factors affect children's sex-role attitudes, aspirations, educational choices, and labor market outcomes.

Another way to examine women's occupational preferences is to look at the household division of labor in families where wives have higher predicted earnings than do husbands. In such a situation, it would make more economic sense for men to drop out of the labor force for childrearing than women. We have identified such families in the PSID and are examining their allocation of time between home and market.

The above research projects are only a beginning in an attempt to sort out the ways

in which sex role socialization influence the supply side of the labor market.<sup>3</sup> As we have said above, we are convinced that empirical work that does not examine the mechanisms whereby men and women both behave differently and are treated differently are unlikely to tell us much that we don't know already. That is, women earn less than equally qualified men do, and this gap is caused either by discrimination or socialization.

<sup>3</sup> We think it is equally important to look at demand-side behavior.

## REFERENCES

- Arrow, K., (1972a) "Models of Job Discrimination," in A. H. Pascal, ed., *Racial Discrimination in Economic Life*, Lexington: D. C. Heath, 1972, 83–102.
- , (1972b) "Some Mathematical Models of Race in the Labor Market," in A. H. Pascal, ed., *Racial Discrimination in Economic Life*, Lexington: D. C. Heath, 1972, 187–204.
- Bergmann, B. R., "Occupational Segregation, Wages and Profits When Employers Discriminate by Race or Sex," *Eastern Economic Journal*, April-July 1974, 1, 103–110.
- Corcoran, M. and Duncan, G. J., "Do Women Deserve to Earn Less than Men?," in G. Duncan et al., *Years of Poverty, Years of Plenty*, Ann Arbor: Institute for Social Research, 1984.
- Eccles, J. P. and Hoffman, L. W., "Sex Differences in Preparation for Occupational Roles," in H. Stevenson and A. Siegel, eds., *Child Development and Social Policy*, forthcoming.
- Marini, M. M. and Brinton, M. C., "Sex-Typing in Occupational Socialization," in B. F. Reskin, ed., *Sex Segregation in the Workplace: Trends, Explanations and Remedies*, Washington: National Academy Press, forthcoming.
- Treiman, D. J. and Hartmann, H. I., *Women, Work, and Wages: Equal Pay for Jobs of Equal Value*, Washington: National Academy Press, 1981.

## Pacific Protagonist—Implications of the Rising Role of the Pacific

By STAFFAN BURENSTAM LINDER\*

Rapid economic growth in the Pacific region is transforming the world. It is interesting to speculate on the multidimensional consequences, concentrating on the economic and political effects.

Different definitions of the "Pacific" may be adopted, depending upon the problem at hand. To direct attention to a new phase in the grand historical process of changes in the economic and political gravity of the world, it may be suggestive to note how the ratio of Pacific *GDP* to Atlantic *GDP* has increased from less than 40 percent in 1960 to almost 60 percent in 1982. In the Pacific Basin is then included the market economies of the Asian-Pacific region plus China and the five Pacific states of the United States. The Atlantic Basin includes the European OECD countries and the eighteen Atlantic states of the United States.

If we, to analyze the mechanics of this process, concentrate on the market economies of the Asian-Pacific region, we can see that, over the same period, this region's share of world *GDP* has doubled and now stands at over 16 percent. This Asian-Pacific group includes Japan—with 60 percent of the region's *GDP*—and the other members of Northeast Asia (South Korea, Taiwan, and Hong Kong), the ASEAN countries (the Philippines, Indonesia, Singapore, Malaysia, Brunei, and Thailand), and, finally, the South-Pacific countries (Australia, New Zealand, and Papua New Guinea).

<sup>†</sup>*Discussants:* Carl E. Beigie, Dominion Securities, Ames; Lawrence B. Krause, The Brookings Institution.

\*Stockholm School of Economics and Parliament of Sweden, S10012 Stockholm, Sweden. I thank the Marianne and Marcus Wallenberg Foundation for support and the Hoover Institution for facilities. Some assistance was given by the World Bank.

The gravity shift is also reflected in trade flows. Asian-Pacific exports as a share of world exports have gone from 9 percent in 1960 to 19 percent in 1983. This rise has been managed in spite of the fact that the region has a small part of oil exports, the enormous value increase of which dwarf many changes in the world trade pattern. If we look at manufactured exports, we can see that of all such exports from *LDCs* as much as 59 percent now comes from the four small Asian-Pacific newly industrialized countries *NICs* (South Korea, Taiwan, Hong Kong, and Singapore). This percentage was 25 percent in 1960. Trade changes can also be highlighted by pointing out that U.S. total trade (exports plus imports) with the Asian-Pacific region has risen from less than 50 percent of U.S. trade with OECD Europe in 1960 to 120 percent now.

In trade, the Asian-Pacific countries already match the United States and Europe. In production and income, they are still considerably smaller with the United States and Europe each representing about 30 percent of world *GDP*. However, with the same growth rate differences in years to come, they would reach the same level of *GDP* as the United States and Europe by the turn of the century. If China, now implementing some growth conducive economic reforms, would approach the Asian-Pacific market countries and attain greater economic weight, the importance of the region would become even more pronounced. But, even with the great increase that has already taken place in the relative importance of the Asian-Pacific group of countries, the outward appearance and the inner nature of the world economy and world politics has changed. This transformation is reflected in many ways.



### I. The Demonstration Effect

To begin, there is the *invigoration effect*. It has been vitalizing for less developed countries and also for public opinion in industrial nations to see that growth can be achieved also by those countries which are behind. There was much development optimism in the period immediately following World War II, and then much disappointment, as the strategies in vogue proved unhelpful. For less developed countries, the takeoff proved much more difficult than first thought. But now, when Asian-Pacific LDCs have proved that it can be done, there is some cause for renewed optimism—granted that the previous ways can be changed.

The Asian-Pacific LDCs have demonstrated that the Japanese miracle is not a miracle as it can be duplicated. There is not just a center and a periphery in some eternal arrangement, as is sometimes concluded from the failures of many. The Asia-Pacific LDCs and their success have shown that to take such a position is to abdicate into self-fulfilling pessimism, easy excuses, and facile accusations. Indeed, as the Asian NICs have demonstrated, it is possible not only to stage vigorous growth but to challenge the old industrial countries and even, as Japan, to reach out for leadership. The Asian-Pacific region radiates confidence and instills confidence.

An important question is, then, through which means has Asian-Pacific growth been achieved? There are many different interpretations. However, the more widespread successful growth has become in the region, the more emphasis has been given the economic explanations. Growth accounts stressing institutional factors like the Japanese ethos, the activities of the Japanese Ministry of International Trade and Industry (MITI), Confucian values, Chinese industriousness, abundant natural resources, or the reverse, that is, a situational imperative in the form of nonexistent natural resources may well be important, but they cannot provide the gist of the story. Their relevance differs widely from one country to another in the Pacific region which, it must not be overlooked, is composed of countries diverse in almost all respects,

All the successful countries in the region have, however, one characteristic in common: they are market economies relying on private entrepreneurship and property. Incentives and efficient allocation are emphasized. Foreign trade is given a prominent place in an outward orientation of the development strategy. In short, these countries are capitalist and not socialist. The socialist countries of the region have met with little success.

It may also be noted that in South Korea and Taiwan, growth did not really start until around 1960 when, in both countries, there was a far-reaching redirection of the economic policies away from regulation and interventionism over to a reliance on market mechanisms.

The success of the Asian-Pacific market economies was unexpected and went for a long time unnoticed, as the attention was concentrated on countries pursuing very different economic policies following the once fashionable growth strategies of planning and inward orientation. Explanations stressing the importance of the market system and capitalist methods in the development of the Asian-Pacific countries have now gained wide currency. This will yield what we may call an *ideological effect*. Political thought and economic thinking will be influenced in the old industrial countries where the faith in the system, which once brought these countries their wealth, is again rising after a long period of erosion. It will affect the non-socialist part the Third World where the market system has mostly been downplayed, or has been in disrepute as market signals have often been so overburdened by the effects of extensive regulations that they have failed to give efficient solutions. Finally, the communist countries will find their situation sharply changed by the success of the Asian-Pacific countries.

There are double problems for the communist countries. Their power has diminished relatively when the economic resources of the Asian-Pacific countries, especially Japan and the NICs, have risen so fast. Secondly, the attractiveness of the communist system, as such, has declined. The communist parties in the Asian-Pacific region languish through the combined result of

the nonperformance of the socialist model and of the achievements under capitalist methods. In spite of a considerable military buildup, Soviet political influence in the region is much smaller than it once was. Indeed, in Asia where, according to Nikita Khrushchev, the world battle for the masses would be won, the position of communism has deteriorated sharply.

The success of the market-oriented countries of the Asian-Pacific region will have an effect not only on thinking but also on acting. The ideological effect will be followed by an *imitation effect* in economic policy. At least in some older industrial countries, market solutions will be tried again more confidently. Among the *LDCs*, there will be some attempts at deregulation. In some of the communist countries, economic reforms will be experimented with because of domestic failures in the socialist East, but will be given direction by the success achieved in the capitalist Far East.

The economic reforms presently undertaken in China are at least to some extent inspired by the demonstration effect of successful economic policies instituted by neighboring countrymen. For China not to slide into giant insignificance, it has been important to try some similar methods. In discussions with people who are active in the Asian-Pacific region, it is easy to get confirmation of the view that there has been and is such a demonstration effect at work. Yet, in the literature on Chinese economic reform, there is no reference to such a demonstration effect. One reason for this is that, given the limited scope for debate in China, there are few or no public references to the examples set by neighbors.

There may well be strong policy reversals in China when some of the reforms unavoidably dilute the central power of the communist party. However, even so the on-going reform activities and the prospect of their success has jolted the position of the Soviet Union as dramatically as the defection of China from the Soviet fold once did. Furthermore, the Chinese reforms in their turn have a strong effect on Third World countries that have found it possible to neglect the economic policies and achievements of the Asian *NICs*. For various reasons these

*NICs* have a weak standing in the politics of the *LDCs*. The politics of Taiwan and Hong Kong are both intimately connected with that of the Peoples Republic of China, Singapore is small and South Korea politically exposed because of its conflict with North Korea. Furthermore, the strong political interests elsewhere, fearful of the ideology effect and dreading the imitation effect, have gone to great lengths to denigrate these market economy superstars. The achievements of the Asian *NICs* are explained away as special cases and attributed a number of negative side effects. These countries have a public relation image far below their performance picture.

## II. A Growth Pole Effect

The Pacific dynamism will make this region serve as a growth pole. There will be a *propagation* effect as set out in international trade theory. The Asian-Pacific countries will be a bigger market, serve as a better source of supplies, and add to international factor movements. The successful trade-oriented development strategy will then benefit not only the Asian-Pacific countries, but also their trading partners. This is an important effect in a sluggish world economy.

It is evident that the trading partners of the Asian-Pacific countries will primarily be those countries which have an open, outward-oriented system themselves, that is, the market economies. It is easy to see in international trade statistics this intensification in exports and imports. Especially for the United States, the proportion of exports which goes to the Asian-Pacific countries has risen fast—from 13 percent in 1960 to 25 percent now.

## III. A Threat Effect?

The trading partners will however be exposed to certain difficulties of transition. Gains from trade and from factor movements are not automatic, as they are assumed to be in the simplest trade models. For the gains to be fully realized, there must be a reallocation of factors of production and, in practice, such a reallocation may be painful to those exposed to the need for it.

The forces of competition are required to bring about a beneficial reallocation. They also give dynamic trade gains, as they stimulate the search for improvements and innovation. Yet, those exposed to these pressures often resist them and seek support in cries for protection. As we know, the authorities sometimes yield to the lobbies and, to the cost of the overall economy, put in place tariffs and nontariff barriers or subsidies to help the ailing sectors. Over the last few years, the protectionist pressures have intensified, the undertakings in international agreements have been increasingly disrespected, and it has been difficult to negotiate mutual concessions. The newcomers—and they are primarily the Asian-Pacific countries—have been blamed for posing a severe threat to the old industrial countries.

To evaluate the *threat effect* of Asian-Pacific growth, we may first draw upon the numerous studies that have shown that competition from Japan, the *NICs*, and the *LDCs* in general has not added as much to the pressures for adjustments as usually argued. Other factors such as technological change, demand shifts, and changes in energy prices have necessitated additional adjustment. Furthermore, a new competitor is a new customer, that is, there is no reason why there should be a net destruction of economic activity or employment opportunities.

Yet, adjustment is required. A new competitor is a new customer, granted that we can supply what the new potential customer demands—this added adjustment burden is the problem. The capacity for adjustment has declined in the old industrial countries at the same time as the need for adjustment has increased. The reallocation requirements caused by the Asian-Pacific countries have been added to a total of pressures which has outsized the reduced ability to cope with change.

There are many reasons why the ability to adapt to new conditions has worsened. Tax policies and social policies have reduced labor market mobility and the incentives to invest in new skills and positions. Labor market legislation has similarly given a preference to what exists, although it cannot be viable forever. Restrictive practices, negotiated and applied by the unions, have made it harder

to move into new activities and technologies. Macroeconomic policies of accommodation and unconditional government undertakings to pursue policies for full employment have reduced the responsibilities of the labor market partners to negotiate contracts that are compatible with short-term competitiveness and long-run employment. Shackles have been substituted for sticks and carrots.

These problems have, in particular, beset the European countries where there have been a rapidly rising unemployment and a decline in industrial employment and in the capacity to generate new jobs. In this predicament, governments have tried to find an easy remedy in introducing various subventionist policies and protectionist measures. In the United States, misaligned exchange rates during the last few years have given an import surge that has provided excuses of a more conventional sort for an escalation of protectionism. The question then is, whether vigorous growth (in the Asian-Pacific region) combined with a reduced-adjustment capacity (particularly in Europe) gives some sort of a new irrefutable protectionist argument—the “geriatric tariff.” The answer is no.

Protectionist measures, whether of the old type or in the form of novel new tariff barriers (*NTBs*), do not eliminate the disadvantages of a low-adjustment capacity. Indeed, they compound them, by introducing even greater obstacles to change. The policies will fail. When the investments of structural change for the long-run good are not undertaken and resources instead are used to live fully in the short run, there will be an accumulation of structural problems. Adjustment pressures will then surface in the form of reductions in income and demand, and in a reduced pull from export industries and industries of the future. Both households and industry will in this way be exposed to pressures, and more difficult pressures, which could otherwise have been avoided. The wealth of the old industrial countries has been created through successful activities on free domestic and international markets and cannot be maintained by unfreeing those markets.

Free trade like charity begins at home. Free international trade is the extension of free domestic trade to take advantage of an

even wider division of labor and competition in an optimization of resource allocation. Unfortunately, for some time a freeing of international trade has been paralleled by a process of rapidly increasing domestic trade obstacles. There is a point when domestic trade is so regulated that free international trade sends the wrong signals through the domestic economy, and this economy propagates the wrong signals through international trade. In this situation the framework of international trade will crumble. It is suggestive to watch the great difficulties of the Comecon countries to enter into fruitful and long-standing international trade relations. Within the bloc there are enormous difficulties with "planned trade." To try to make different plans meet means to multiply problems that are already unmanageably big. Trade with market countries is made uneasy as the market countries are not prepared to accord most-favored-nation treatment and free-trade conditions to countries where competitiveness is not determined by efficiency, but by government decree and can be changed from one day to another.

When non-socialist countries now devise systems for planned trade and resort to the potentially even more distorting method of subsidizing exports, it is important not to forget these problems. Regulations and subsidies, and measures to countervail the trade effects of regulations and subsidies, are wearing thin the fabric of international trade.

As the theory of comparative costs teaches us, it is not possible for anyone, no matter how vigorous he is, to outcompete others all around, as is sometimes alleged. However, those with a low adaptability will see the gains from producing and marketing innovative products on a world market scale taken by others. Even with low flexibility, however, income levels on average would be higher than those with restricted trade, but would be far below what they could be with more vitality and initiative.

The strength of Asian-Pacific competition does not make refuted protectionist arguments relevant. That the new actors are highly vigorous does not change anything except that it raises the benefits to be reaped by powerful interaction and that it increases the costs of a failure to respond constructively.

Yet, the threat effect of Asian-Pacific growth is very real, even if its nature is not what its proponents suggest. The threat is that the old industrial countries have lost the flexibility needed to grasp opportunities and that they, in the process of trying to preserve what they have created, ruin it by permitting under competitive pressures, destructive policies to multiply. The menace is not the vitality of some, but the sclerosis of others. As argued before, there is an imitation effect of Asian-Pacific growth in that some try to emulate. However, the threat effect arises in that others, rather than imitating what has succeeded, adopt damaging policies.

The specter of the "Yellow Peril" is again raised with the suggestion that, somehow, there could be an inundation rather than a two-way flow of trade and that there is a permissible excuse for raising trade barriers. But, instead of a Yellow Peril, there is a "White Peril." The White Peril arises from a gradual increase in barriers to domestic trade, and from an extension of foreign trade barriers, which together lead to an accumulation of structural problems. The barriers will harm the old industrial countries and endanger the world trading system. The old industrial countries find it convenient to rationalize their new trade barriers by referring to trade obstacles operated by the Asian-Pacific countries.

Market-based Asian-Pacific growth upsets the confidence of the socialist countries in their system and thus creates several political problems for these countries. It ought to have a positive effect on the beliefs of the Western market economies. In fact, there are political problems for the Western countries, too. Exceptionally high growth in the Asian-Pacific region will unavoidably cause a decline in the relative weight not only of communist countries, but also of the Western market economies. However, this decline will be even more pronounced to the extent the faltering response to the Asian-Pacific challenge (i.e., the White Peril) will strain the political system and the social harmony in the old industrial countries. An expanding battery of obstacles to domestic and foreign trade would aggravate the performance. Second, it will put sticks in the spokes of the wheels of export-led growth in the friendly,

and potentially even more friendly, Asian-Pacific region. There is the possibility of these countries being antagonized by being the victims of discriminatory protectionism and of accusations levied by those who use them as scape goats. The political costs of this are considerable for the Western countries that, instead, should be able to gain handsomely from improved relations with these important partners for the future.

Third, a negative Western response in trade policy will scare off those developing countries that are on the verge of imitating the Asian NICs but now find it seemingly hazardous. The former trade pessimism inspired by the development thinking of the 1950's and 1960's has subsided, but only to be replaced by a new export pessimism caused by the widespread protectionism of the 1980's in the old industrial countries. This pessimism—like the original pessimism

—is overdone as, after all, it has proved possible to expand exports considerably. Yet, the new mood has a definite impact on those countries that may contemplate whether they should move from an inward-looking to an outward-looking strategy. This would be damaging, as it would bar the best or only avenue to material improvement. For the West it is also a political problem, as it would push countries away from applying Western economics methods. They would not be attracted into a more cooperative and less antagonistic Third World.

Asian-Pacific dynamism has important implications. The basic message is this: the newcomers provide enormous new opportunities but threaten, politically and economically, those who in the First, Second, or Third Worlds permit political or economic rigidity to dominate the response.

# Is There Need for Economic Leadership?: Japanese or U.S.?

By W. W. ROSTOW\*

There are two reasonably unambiguous definitions of the inherently ambiguous concept of economic leadership. Definition One relates to innovation and leading sectors; that is, the relative primacy of a country in commercializing a new technology and establishing, for a time, a dominant position in a major sector. In that sense, Britain led in the first phase of the cotton textile revolution (say, 1783–1832), and the United States led in the first phase of the mass automobile revolution (say, 1909–29).

Definition Two relates to policy; that is, the assumption of responsibility for the successful operation of the world economy as a whole by a single country. I shall consider later the necessary conditions for such leadership.

The leading sector and policy definitions are partially linked because only an economic power quick off the mark in converting new inventions into profitable innovations is likely to be able to sustain the balance of payments implications of policies of responsible leadership in the world economy.

I shall now respond to the subject of this paper in terms of each definition of leadership. First, the new technologies and their implications for leadership in the innovational, leading-sector sense. By new technologies I refer to the microchip, genetic engineering, the laser, robots, new communication methods, and new industrial materials. Although germinating for some time—and by no means uniform in their timing—I believe historians will date the innovational stage of this technological revolution from, roughly the second half of the 1970's.

Somewhat arbitrarily, I am inclined to regard this rather dramatic batch of innovations as the fourth such major grouping in

the past two centuries.<sup>1</sup> The fourth industrial revolution has some distinctive characteristics as compared to its predecessors. It is more intimately linked to areas of basic science which are themselves undergoing rapid revolutionary change. This means the scientist has become a critical actor in the drama; and the successful linkage of the scientist, engineer, and the entrepreneur has become crucial to the generation and diffusion of the new technologies. The new technologies are also proving ubiquitous, progressively suffusing the older basic industries, agriculture, animal husbandry, and forestry, as well as all manner of services from education and medicine to banking and communications; and they are, in different degree, immediately relevant to the economies of the developing regions, depending on their stage of growth, absorptive capacity, and resource endowments.

For our purposes, the extraordinary range and diversity of the new technologies bear directly on the prospects for leadership by Japan, the United States, or anyone else. I find it most improbable that any one nation will achieve and sustain across-the-board technological leadership in the fourth industrial revolution, or, indeed, leadership in

<sup>1</sup>The first industrial revolution, dated by innovation rather than invention, came on stage in the 1780's and embraced factory-manufactured cotton textiles, good iron made from coke, and Watt's more efficient steam engine. The second starts in the 1830's and becomes an extremely large-scale enterprise in Great Britain and the American Northeast in the 1840's; i.e., the railroad which, within a generation, induced the invention of cheap mass-produced steel. The third comes round about the turn of the century and consists of electricity, the internal combustion engine, and a new batch of chemicals. In their various elaborations they run down to the second half of the 1960's, when the leading sectors of the third industrial revolution decelerate markedly. For further discussion, see my book (1983a especially pp. 54–60; 88–94). Also see my lecture (1984, especially pp. 3–10).

\*Professor of Political Economy, University of Texas, Austin, TX 78712.

a major area such as micro electronics or genetic engineering or new industrial materials. Each such area represents, in fact, a group of highly specialized and differentiated activities. Given the reasonably even distribution of scientific, engineering, and entrepreneurial talent among the advanced industrial countries—and the similar educational level and skills of their working forces—with the passage of time, specialized comparative advantage is likely to be distributed among a considerable range of countries; and one is likely to see a great deal of cooperation and trade in the new technologies, as well as competition. Indeed, if one examines the pattern of joint ventures across international boundaries and the expanding trade in high-technology sectors, the process can already be seen to be under way, despite the somewhat slower start of Western Europe than Japan and the United States in exploiting the new possibilities.

The diffusion of virtuosity in the new technologies will be accelerated by their indirect impact on the developing regions. Over the next decade we are likely to see the new technologies vigorously applied in the motor vehicle, machine tool, steel, textile, and other industries rooted in the longer past. One result of this conversion to high tech along a broad front is that the more advanced developing countries will no longer be able to count on generating increased manufactured exports simply by exploiting their lower money-wage rates. There is a lively awareness of this change in prospects in the Pacific Basin because of palpable Japanese progress in applying the new technologies to the older industries. In consequence, there is intense interest among the newly industrialized countries in acquiring the emerging technologies. The Republic of Korea, for example, is gearing its current Five-Year Plan to the rapid absorption of the new technologies, including quite radical changes in education policy. (For an extended discussion, see my 1983b book.)

Each developing country differs, of course, in both the extent to which the new technologies are relevant and in its capacity productively to absorb them. But, in general,

potential absorptive capacity is higher than relative per capita levels of real income would suggest.

Consider the case of India, a country with an exceedingly low average real income per capita, conventionally measured. The pool of scientists and engineers in India has increased from about 190,000 in 1960 to 2.4 million in 1984 (see Government of India, 1984); and it is sustained by the fact that something like 9 percent of the Indian population aged 20–24 is now enrolled in higher education, three times the proportion twenty years earlier. Taken along with the large absolute size of India's population, this means that India is quite capable of assembling the critical mass of scientists and *R&D* engineers required to solve the kinds of problems increasingly posed by the fourth industrial revolution and its efficient absorption.

The central question is whether Indian society can achieve, over a wide spectrum of sectors, the bringing together in partnership of scientists, engineers, and entrepreneurs which has happened in atomic energy, space, and, to a significant degree, in agriculture. The linkage has not been effectively made in a good many industrial sectors. I would guess that if these linkages began to firm up in one sector after another, the Indian brain drain would begin to reverse.

While the Indian case is rather dramatic, given the country's size and relatively low real income level, the revolutionary expansion of higher education over the past generation is quite typical of the developing world. For middle-income economies as a whole, the higher education proportion of the relevant age group rose from 3 percent in 1960 to 11 percent in 1979. (See World Bank, 1983.) It will certainly take time for these more advanced developing countries to bring about the partnership of scientists, engineers, and entrepreneurs the absorption of the Fourth Industrial Revolution requires. (It is, indeed, taking quite a lot of time in Western Europe and the United States.) But I do believe it will happen; and the process will strengthen the diffusion of technological virtuosity within the world economy.

Now, what about the second definition of leadership, in terms of policies reflecting responsibility for the viability of the world economy? The capacity to lead, in this sense, depends in part, of course, on a nation's proportionate role in the world economy; its relative contribution to global *GNP* and international trade; and the strength of its capital markets and the scale of its international lending. Britain's contribution to global industrial production may have fallen from 32 to 14 percent between 1870 and 1913, its foreign trade from 25 to 16 percent (see my 1978 book); but its large and active capital market, combined with the maintenance of a free-trade policy, permitted Britain to remain an acknowledged leader in the world economy down to World War I.

There was, however, more to it than that. British leadership was only possible because the United States, the major states of Europe, and the component regions of the British Empire by and large conducted their business in ways compatible with London's rules of the game. They did not generally share London's passion for free trade; but that fact did not prevent the relatively easy flow of goods and capital and people and the acceptance of transmission mechanisms which kept the world economy roughly in step with respect to prices and cyclical fluctuations.

At the base of the system was a fundamental, shared agreement in domestic politics; namely, that tariffs apart, there was no realistic alternative to accepting the domestic consequences of the vagaries inherent in a competitive, largely private enterprise global system of trade and capital movements.

The tragic experiences of the world economy from 1920 to 1939 radically altered both the international and domestic aspects of the pre-1914 consensus. At the close of World War II, the United States, conscious of its interwar derelictions and of its extraordinary relative economic strength in a war-damaged world, explicitly accepted responsibilities for economic leadership, including initially high levels of official grants and loans to foreign governments.

The U.S. role was sustained by a widespread consensus on appropriate internation-

al and domestic rules of the game, the latter including the acceptance by governments of responsibility for both the domestic level of employment and the expansion of social services.

Over the past forty years, the relative role of the United States in the world economy has declined with the revival of Europe, West, and East; with the extraordinary surge of Japanese growth; and with the expansion of the developing economies at overall rates averaging higher than those in the advanced industrial economies. As of 1980, the United States contributed perhaps 23 percent to global *GNP*, the figure having declined from about 33 percent in 1950. (The U.S. and Japanese *GNP* data in relation to global *GNP* are from Herbert Block, 1981, pp. 30-32, Appendix Table 1.) The Japanese *GNP* proportion rose between 1950 and 1980 from about 3 to 8.5 percent.

Despite the sharp relative rise of Japan, the capacity of the United States to lead, in terms of my second definition, is evidently still greater than that of Japan; and that capacity is enhanced, as was that of pre-1914 Britain, by the continuing large role of the United States as a capital market. But the fact is that, no more than Edwardian Britain, does the United States command the power to impose its leadership on the world economy. The distribution of effective economic and political power is, in fact, greater now than then.

My interim conclusions, then, are these. First, whether one uses leading sector Definition One, or rules of the game Definition Two, forces are at work tending to diffuse, rather than concentrate, the power to lead. Second, an overwhelming concentration of power in a single country does not appear to be a necessary condition for effective leadership in the world economy, if there is an effective working consensus on economic rules of the game, domestic as well as international. It is here, I believe, in domestic policy that the critical problem lies which led us to focus this session on the question of leadership in the world economy.

The circumstances that emerged in the world economy in the early 1970's broke up



the consensus on domestic economic policy that had emerged after World War II and acquired legitimacy with the historically unique growth rates of the 1950's and 1960's. No viable successor consensus on domestic economic policy has yet emerged. And it is the lack of an effective consensus on domestic policy in Western Europe and the United States which mainly accounts for the rise of protectionism, distorted real interest rates, precarious debt structures, a grossly overvalued dollar, and other pathological aspects of a disheveled world economy. Specifically, international economic policy is likely to remain ineffectual until the West learns how to reconcile relatively full employment (and reasonably high and steady growth rates) with control over inflation, and how to generate and absorb rapidly the technological possibilities inherent in the fourth industrial revolution.

This judgment stems from a particular view of where the Atlantic world stands in the sweep of modern history. For a century now—from, say, Bismarck's first major social legislation in 1883—the central problem addressed in the advanced industrial nations of Western Europe and North America has been, how can we build industrial societies which reconcile efficiency in a world of rapidly evolving technologies with the humane values in which Western culture is rooted? In politics the process often assumed the form of debate and struggle within a zero-sum game which allocated resources as between welfare and private consumption and investment, as between the less affluent and more affluent. It is one of the major achievements of the Western democratic process that this muted form of class struggle proceeded in relative peace, reaching in the decades after World War II a remarkable apogee. The proportion of social outlays rose in seven major OECD countries from 14 to 24 percent of *GDP* between 1960 and 1980 (see *OECD Observer*, 1984)—a truly revolutionary shift.

Since trees do not grow to the sky, the expansion of social outlays at rates higher even than the extraordinary real growth rates of the 1950's and 1960's was bound, in time, to cease. Pressures to contain these outlays

increased sharply with the explosion of oil prices, exacerbated inflation, and high unemployment rates of 1974–80. After three years of remission, a similar traumatic sequence occurred in 1979–80. This time, the recession continued for a further two years, in part due to U.S. domestic economic policy. The *GDP* per capita, which has grown at 3.8 percent per annum from 1950 to 1973 for the advanced industrial countries, decelerated to 2.0 percent for the period 1973–79, and averaged slightly negative over the next three years. Meanwhile, amidst these setbacks the fourth industrial revolution asserted itself on the world scene with its potentialities, challenges, and dangers for those who lagged in its exploitation.

Clearly, the central issue of domestic political life in the West could no longer be defined in terms of the allocation of a bit more or less to social welfare. Two issues, above others, appeared fundamental if the erosion of the social and physical infrastructure of the advanced industrial countries was to be avoided: the mounting of long-term policies capable of reconciling relatively high rates of growth with control over inflation; and the bringing together of scientists, engineers, entrepreneurs, and the working force to generate and absorb efficiently—and across the board—the technologies of the fourth industrial revolution. In my view at least, the former objective requires effective long-term incomes policies, as well as a judicious blend of fiscal and monetary policies. Evidently, incomes policies are essentially political, social, and institutional, rather than narrowly economic arrangements; and much the same can be said of the partnerships rendered imperative by the peculiar character of the fourth industrial revolution.

Looked at in this way, the advanced industrial countries of the Atlantic world are caught up in a transition between the struggle over the allocation of resources that marked the evolution of the welfare state, and the need for sustained communal cooperation if growth and control over inflation are to be reconciled and the new technologies find an appropriate role in their societies. If one pierces the veil of political rhetoric—which is exceedingly slow to change—

one can observe the transition as a halting de facto process in Western Europe and in the United States.

Assuming this view has a reasonable degree of validity, it throws some light on our instinctive feeling that, somehow, Japan is, at the moment, better geared to the tasks of the generation ahead than the major countries of the Atlantic world. From, say, the first ineffectual Factory Act of 1898 (which aimed to protect women and children) down to about 1936, Japanese social legislation followed, with a lag, a sequence not unlike that to be observed in the Atlantic nations at similar stages of growth. (For a brief summary of this sequence, see my 1971 study, pp. 149–52.) But a generation of war, physical destruction, and postwar recovery broke the pattern and forced Japan into a sequence of communal efforts climaxed by its extraordinary surge of growth since the mid-1950's. A lively sense of the communal stake in the continuity of high growth rates helps account for the fact that Japan has, in a sense, already made the transition towards which much of the Atlantic world is moving rather slowly and with some pain; that is, Japan enjoys a quite well-institutionalized incomes policy and has adjusted its institutions with alacrity to the imperatives of the fourth industrial revolution. There are within Japan debates and struggles about the appropriate proportionate level of social outlays and about income distribution, and these could become more acute; but there is also a more solid consensus than in the Atlantic world that these zero-sum contests are less fundamental than the common effort to assure a steadily expanding pie to be appropriately divided.

These qualities render Japan an extraordinarily strong unit in the world economy; but its capacity to lead is restricted by its incomplete acceptance of the trade responsibilities that leadership demands, the still limited, if expanding, capacity of its international capital markets, and, above all, by the lack of consensus on domestic economic policy in the rest of the OECD world.

I conclude, then, that forces at work in the world economy are likely to diffuse, rather than concentrate, leadership over the time

ahead under either of the two definitions with which I began; that such dilute multi-lateral leadership is quite feasible if a reasonably wide consensus exists on the economic rules of the game, both domestic and international; that the United States is in a position to provide, potentially, a higher degree of leadership than Japan; but it is gravely inhibited because it is still caught up in a transition to a new post-welfare-state consensus on domestic economic policy the Japanese have substantially made.

To answer directly the question posed in the title of this paper—the world economy needs the leadership of both Japan and the United States, but, for different reasons, neither is in a position to supply it at the moment and neither can do the job alone.

The extent to which the United States can provide its share of leadership, over the next several generations, will depend substantially on how fast we diffuse the new technologies across the old basic industry sectors as well as agriculture and the services. I emphasize diffusion rather than generation because, by historical accident—including the land grant colleges and the commitment to serve the economy in which they are rooted—we are almost certainly the best positioned of all the advanced industrial countries to build the close, flexible working partnership of scientists, engineers, and entrepreneurs necessary to create and innovate the new technologies. Indeed, we are in tolerably good shape in three major sectors which arose from laboratories and have managed to maintain the interactive osmotic partnership between scientists and the production process necessary for competitive viability; that is, electronics, chemicals, and aerospace. But our entrepreneurs in steel, motor vehicles, machine tools, and some other basic industries have not worked well with the *R&D* process in the post-1945 years.

I might add that our problem has been compounded by business schools that teach their students how to maximize the bottom line with fixed production functions, but not how to operate in a world of rapidly changing technological possibilities and accelerated obsolescence. And we economists haven't helped much. For more than two

centuries we have failed to build the process of generating and diffusing technologies into the mainstream propositions of macro and micro theory.

Nevertheless, I would guess that with costly time lags, the United States will solve the critical problem of diffusion and meet rather well the challenge of the fourth industrial revolution.

But there is a lion in our path; that is, the four interlocked pathological problems that now have us in their grip: excessively high interest rates, a grossly overvalued dollar, a scandalous balance of payments deficit, and a federal budget deficit running at 5 percent of *GNP*. Sooner or later we shall have to deal with these problems that we are now pushing down the road by borrowing on a profligate scale. The question is whether we can move to a more viable balance without triggering a grave national and international economic crisis.

This is clearly not an occasion to offer detailed policy prescriptions. I would only observe that it seems most unlikely that those now deeply rooted problems will be resolved merely by a somewhat easier monetary policy, a somewhat tighter fiscal policy. I believe at some stage a rigorous incomes policy will be required, not merely to contain the inflationary effects of the necessary dollar exchange rate adjustment, but also to render credible the decline in interest rates and to assist in narrowing the federal deficit.

We have been living grossly beyond our means and we won't get our books into balance merely by a bit of jiggery-pokery with macroeconomic policy.

But I doubt that American political life has the capacity to face these problems until we are in much worse trouble than we are right now. We are more likely to operate on the principle Jean Monnet set out in his *Memoirs*: "...people only accept change when they are faced with necessity, and only recognize necessity when a crisis is upon them" (1978, p. 109).

As an economic historian, I have long rejected the convenient but illusory theoretical distinction between the long run and the short run. The Marshallian long run is moving every day of our lives. Put another way,

the long run is simply the accumulation of what happens over short periods of time. Therefore, there is a clash between my temperate optimism about the United States and the challenge of the new technologies, and my temperate pessimism about our capacity to correct the distortions in our economy without major crisis. A crisis is likely to slow up, for a time, the pace at which we carry forward the fourth industrial revolution.

But looking ahead over the 1980's and 1990's, my net judgment, for what it may be worth, is that Mancur Olson hasn't gotten us yet, that we remain a resilient continental society, with many changing centers of energy and initiative, and that we will sustain a vital role in the economic life of the world economy, as the inevitable and wholesome diffusion of economic power continues as it has since the late 1940's.

But is the diffusion of economic power consistent with leadership and a reasonable degree of order? Or, are we on the road to perpetual chaos in the world economy? On this matter I would supplement my earlier observation on the importance of consensus concerning domestic rules of the game with an institutional recommendation. We are most likely to master chaos in a world of diffuse power by going to work, in the first instance, regionally rather than globally; that the Pacific Basin is a promising arena for demonstration; and the place to start is with the Djakarta initiative of ASEAN of July 12, 1984.

## REFERENCES

- Block, Herbert, *The Planetary Product in 1980: A Creative Pause?*, Bureau of Public Affairs, Department of State, Washington: USGPO, 1981.
- Monnet, Jean, *Memoirs*, Garden City: Doubleday, 1978.
- Rostow, W. W., *Politics and the Stages of Growth*, Cambridge: University of Cambridge Press, 1971.
- \_\_\_\_\_, *The World Economy: History and Prospect*, Austin: University of Texas Press, 1978.
- \_\_\_\_\_, (1983a) *The Barbaric Counter-Revo-*

- lution: Cause and Cure*, Austin: University of Texas Press, 1983.
- \_\_\_\_\_, (1983b) *Korea and the Fourth Industrial Revolution*, Seoul: Korean Economic Research Institute, 1983.
- \_\_\_\_\_, *India and the Fourth Industrial Revolution*, (Dr. Vikram Sarabhai Memorial Lecture), Ahmedabad: Institute of Management, 1984.
- Government of India**, *A High Tech India Needs You to Take the Right STEP* (The Science and Technology Entrepreneur's Park), New Delhi: Department of Science and Technology, 1984.
- OECD**, "Social Expenditures: Erosion or Evolution?" *OECD Observer*, No. 126, January 1984, 3-6.
- World Bank**, *World Development Report 1983*, New York: Oxford University Press, 1983, Table 25, pp. 196-97.

# THE THEORY OF ECONOMIC ORGANIZATIONS<sup>†</sup>

## Human Fallibility and Economic Organization

By RAAJ KUMAR SAH AND JOSEPH E. STIGLITZ\*

Doctrines concerning what is a good way to organize a society have influenced human societies more deeply than any other set of doctrines. Specifically, beliefs that one way of organizing production and exchange is better than others have inspired a number of socioeconomic experiments leading to modern capitalist and socialist societies, with far reaching implications. Yet surprisingly, the central doctrines, though a source of continuing ideological debate, have been the subject of only limited scientific enquiry. The major proposition, the so-called Lange-Lerner-Taylor Theorem asserting the equivalence between competitive capitalist economies and decentralized socialist economies which make use of the price system, made a point in stressing that the issue of the ownership of the means of production might not be central in the comparison of economic systems. The theorem, however, was based on models both of capitalism and of market socialism in which the most important differences between the two systems were suppressed.

This paper describes a research program attempting to delineate some of the critical differences among alternative forms of economic organization. We contend that central to an understanding of these differences is an understanding of differences in the organization of decision making; of who gathers what information, how it gets communicated and to whom, and how decisions get made, both concerning what actions to take and who should fill decision-making positions. This

view should be contrasted with the traditional economic paradigm in which decision making plays no role: the manager, for instance, simply looks up in a book of blueprints what the appropriate technique of production is for the given set of factor prices. In the conventional paradigm, moreover, mistakes are never made, either in gathering or transmitting information, or in making decisions, and indeed, there are no costs associated with these activities. By contrast, the view we take here is that "to err is human," and that different organizational systems differ not only in what kinds of errors individuals make in them, but also in how the systems "aggregate" errors. As a result, organizations differ systematically in the kinds of errors they make, and thus in their overall economic performance. Organizations also differ in the costs associated with information collection, with information communication and processing, and with decision making. Indeed, as we discuss below, perfect decision making can be achieved by arranging enough decision makers in an appropriate manner, no matter how fallible the decision makers are, provided their decisions are not purely random (or worse). What prevents perfect decision making is the cost.

We refer to the specification of the structure of information gathering, communication and decision making as an organization's *architecture*. The objective of our research program has been to construct stylized models of an economic organization within which the consequences of alternative organizational architectures can be examined.

Using these models, not only can we compare the performance of particular organizational forms but we can also ask, given a particular set of objectives and circumstances, what is the optimal structure (within

<sup>†</sup>*Discussants:* David Kreps, Stanford University; Paul Milgrom, Yale University.

\*Yale University, New Haven, CT 06520, and Princeton University, Princeton, NJ 08544, respectively. This is a shortened version of our working paper (1984a).

a class of structures); for example, what is the optimal number of levels in a hierarchy. We have constructed both a positive and a normative theory. Thus our approach allows us to assess the validity of many of the traditional claims concerning the merits of alternative systems.

### I. The Basic Model

To illustrate the basic principles at issue, we present the simplest possible formulation. Consider an organization facing the problem of choosing among a large number of available projects that are of two types: good projects with an expected return of  $x_1$  and bad projects with an expected return of  $-x_2$ . A fraction  $\alpha$  of the projects are good. With perfect information, all good projects, and no bad projects, would be undertaken. A decision maker makes a judgment about whether the project is good or bad based on whatever information he has. We assume that the information available to one decision maker cannot be fully communicated to others. Here we represent a polar form of this "limited" communication such that individuals within any organization convey to one another only whether they judge a project to be good or bad, even though they might have more information at their disposal.

All decision making is imperfect; we assume that the probability that a decision maker judges a good project to be good is  $p_1 < 1$ , and that the probability that he judges a bad project to be good is  $p_2 > 0$ ; the fact that there is some filtering is reflected in  $p_2 < p_1$ .

We consider two different organizations, each consisting of two individuals. In a *polyarchy*, each individual has the right to accept a project and projects rejected by one are evaluated by the other. Thus, the probability that a good project is accepted is  $p_1$  (the probability of acceptance in the first evaluation) plus  $(1 - p_1)$  (the probability of rejection in the first evaluation) times  $p_1$ , the probability of acceptance in the second review. The probability that a good project is accepted is thus:  $f_1^P = p_1(2 - p_1)$ . Similarly, the probability that a bad project is accepted

is:  $f_2^P = p_2(2 - p_2)$ . The (expected value of the) output of the organization is  $Y^P = \alpha x_1 f_1^P - (1 - \alpha)x_2 f_2^P$ .

By contrast, in a *hierarchy*, for a project to be undertaken, it must be approved by both levels of hierarchy; the probability of this for a good project is  $f_1^H = p_1^2$ , and for a bad project it is  $f_2^H = p_2^2$ . The output of the hierarchical organization is  $Y^H = \alpha x_1 f_1^H - (1 - \alpha)x_2 f_2^H$ .

Two results immediately emerge: *polyarchical organizations accept more bad projects* ( $f_2^P > f_2^H$ ); while *hierarchical organizations reject more good projects* ( $f_1^H < f_1^P$ ). The fact that the two systems make different kinds of errors suggests that there will be circumstances under which one or the other performs better. We can ascertain those conditions by comparing the net output of the two systems:

$$(1) \quad Y^P \geq Y^H$$

$$\text{as} \quad \alpha x_1 p_1(1 - p_1) \geq (1 - \alpha)x_2 p_2(1 - p_2).$$

Condition (1) has a natural interpretation. Assume that we conducted two tests of the project simultaneously; projects with both tests turning out positive should clearly be accepted, those with both tests turning out negative should be rejected (otherwise, there would be no reason to run the tests). The question arises what should we do in a split decision. The probability of a split decision for a good project is  $2p_1(1 - p_1)$  and for a bad project it is  $2p_2(1 - p_2)$ . Thus, the expected value of projects with a split decision is  $2[\alpha x_1 p_1(1 - p_1) - (1 - \alpha)x_2 p_2(1 - p_2)]$ . We accept projects with split decisions if their expected value is positive, and reject them if this is negative. But this is precisely the condition (1); if a project would have been accepted with a split decision, there is no point in having a second review; polyarchy is preferable to hierarchy. On the other hand, if a project with a split decision would have been rejected, then the second review is of value: hierarchy is preferred to polyarchy. Second reviews are also of greater value, in this context, when the ratio of bad projects to good projects is larger and when the losses from bad projects relative to the gains from good projects are greater.

There is another interpretation of our result, which becomes clearest when  $\alpha x_1 = (1 - \alpha)x_2$ . Then the condition (1) is equivalent to the condition  $p_2 \leq 1 - p_1$  where  $p_2$  is the probability of accepting a bad project (Type II error),  $1 - p_1$  is the probability of rejecting a good project (Type I error). In this central case, whether a polyarchy is better than a hierarchy depends on the relative likelihood of the decision maker making the two types of errors. (A more general formulation of this problem is examined in our 1984a paper.)

**Committees:** Another type of organizational architecture is a committee. Committees are collections of individuals with well defined rules for decision making (adoption of a project); for example, majority rule or complete unanimity. In a committee with  $n$  members which requires the approval of at least  $k$  members for adoption, the probability of acceptance of a project, for which the individual's probability of a favorable review is  $p$ , is  $f^c(n, k, p) = \sum_{j=k}^n \binom{n}{j} p^j (1-p)^{n-j}$ .

There is a formal similarity between a polyarchy and a committee in which only one individual needs to approve a project for it to be undertaken, and between a hierarchy and a committee in which all individuals in the committee need to approve the project. In our 1984b paper, we have examined the optimal size and decision rules for committees, and show, for instance, that for fixed  $n$ , output is a single peaked function of  $k$ ; and that, for the symmetric case ( $\alpha x_1 = (1 - \alpha)x_2$ ), whether more or less than a majority consensus should be required depends simply on whether  $p_2 \geq 1 - p_1$ .

## II. Complex Organizations

It should be apparent that most organizations are not one of the pure forms (hierarchy, polyarchy, or committee) discussed above, but rather a mixture of organizational forms. We can build more complex organizations out of these basic building blocks. Consider, for instance, a polyarchy of hierarchies in which to be approved, a project must receive the approval of at least one of several decision units; but each decision unit is, itself, a

hierarchy. This corresponds, rather loosely, to a market economy. If each unit has  $n$  hierarchical layers, and there are  $m$  such units in a polyarchy, the probability of a type  $i$  project being accepted is  $f(n, m, p_i) = 1 - (1 - p_i^n)^m$ . By an appropriate choice of  $n$  and  $m$ , we can ensure that  $f(n, m, p_1) > p_1$  and  $f(n, m, p_2) < p_2$ ; that is, the  $(n, m)$  polyarchy-hierarchy accepts more good projects and rejects more bad projects than a single decision maker. Moreover it rejects more bad projects than the simple polyarchy (for which  $n=1$ ) and accepts more good projects than a simple hierarchy (for which  $m=1$ ). We can construct a more complex organization in which the above polyarchy of hierarchies serves as a single unit, and a second level polyarchy of hierarchies is created using such units. With enough such levels to our complex organization, we obtain perfect screening. It is the cost of decision making that prevents perfect decision making. The cost of a multilevel organization rises rapidly with the number of levels, while there is diminishing returns to the increase in expected output. The balance between these two yields an optimal structure to the organization.

The intuition behind why, with enough layers, perfect screening can be obtained is simple. A hierarchy is successful in increasing the proportion of good to bad projects, but only at the expense of throwing out a lot of good projects. In a polyarchy, on the other hand, the stock of projects under consideration is "replenished." In fact, as the number of units in a polyarchy increases, the probability of a project, of any type, getting accepted goes to one. The sense of our result is that if we first filter the set of projects through a hierarchy, and then through a polyarchy, the set of approved projects is more "refined" than what could have been obtained by a single filter; repetition of such a procedure yields perfect selection.

We can use the analysis of complex organizations in the preceding section to prove other interesting results: for instance, if the organizational cost depends only on the number of managers, then *it is better to reorganize a very long hierarchy into two (or*

more) *polyarchies*: and it is better to rearrange a very large polyarchy as two (or more) *polyarchic subunits* within a hierarchy. (See our 1984b paper.)

### III. Selecting Managers: Towards Organizational Dynamics

Among the most important decisions made within any organization are those concerning who will fill what jobs. The fact that such attention is focused on this problem suggests that it makes a difference; individuals differ in their abilities to acquire, communicate, and process information. For each organizational architecture, there is an optimal assignment of individuals with different characteristics. More importantly, the performance of some organizational forms is more sensitive than others to how these assignments are made, and the errors in assignment are themselves functions of the organizational architecture.

The "rules" by which an organization chooses its successors give rise to stochastic processes, describing the assignments of individuals with different abilities to different positions within the organization. We can analyze, say, the steady state of these stochastic processes for various organizational architectures, and compare their relative performance. We illustrate here how this may be done.

Assume that there are two types of individuals, competent (*C*) and incompetent (*I*). Further, suppose that each person in a polyarchy chooses his own successor, whereas the higher level hierarch chooses his own successor as well as that of his subordinate. Clearly, there are four states of a system: (*C, C*), (*C, I*), (*I, C*), and (*I, I*). We can show  $Pr(C, I)$  and  $Pr(I, C)$  are larger in a polyarchy, whereas  $Pr(C, C)$  and  $Pr(I, I)$  are larger in a hierarchy. If the average output of the two systems were the same, any risk averse society would prefer a polyarchy. (For details, see our 1984c paper.)

Alternative economic organizations also differ in their ability to correct selection mistakes. We suspect, therefore, that a still stronger case for polyarchy can be made

once (evolutionary) mechanisms for eliminating deficient organizations are introduced (for example, bankruptcies in a market system).

### IV. Extensions

The formulation presented here ignores three aspects of cost determination associated with organizational design: (a) time (the more levels to an organization, the greater the time required for decision making); (b) communication costs (not only the direct costs of communication, but also the errors which inevitably arise in the process of communication); and (c) the sequence of decision making (for instance, all individuals are assumed to review all projects in a committee, whereas the upper levels in a hierarchy review only those projects which have been sent up to them by lower levels). Since there are costs to review, the sequencing of the review process may have considerable effect on the resources spent on evaluation.

Elsewhere, we have explored these and other extensions of the basic analysis including (a) a more extensive treatment of the consequences of the use of Bayesian decision rules (though, given the information technology of our basic model, our analysis is Bayesian); (b) an analysis of the endogenous determination of the level of expenditures on information acquisition; (c) an analysis of the consequences of alternative organizational forms on the set of available projects (the mix of projects available to an organization itself is an endogenous variable, determined by the incentives provided to those who develop projects; this in turn is partly dependent on the likelihood of projects of different types being adopted, which differ markedly across organizational architectures); (d) an evaluation of alternative organizational forms faced with different problems, for example, choosing the best set of projects, rather than maximizing the expected profit; and (e) externalities. For instance, one organization's decision affects the productivity of projects undertaken by other organizations. Such interactions are important in certain circumstances, and they may strengthen the case for hierarchy.



### V. Concluding Remarks

The differences in the nature of the errors made by different organizational architectures, though important, are not the only differences among organizational forms. We have not examined the widespread belief of a correspondence between economic architecture and political structures; for instance, the alleged correlation between hierarchies and authoritarianism. We have ignored some aspects of organizational comparisons which have already received extensive discussion in the literature. For example, the traditional models emphasize the computational advantages of decentralization (indeed, in some versions, the differences in organizational forms appear to be simply a comparison between alternative algorithms for solving a general equilibrium problem). The fact that the economy solves the problem in "real time" while the models solve for the equilibrium in pseudo time may mean that (at least as traditionally presented) this argument is of only limited relevance.

We have also ignored the problems of incentives, which have been so much at the center of recent discussions of organizational design. We believe these considerations are not only important, but also that the organizational structure may have a significant effect in determining the set of feasible incentives. For instance, when there is more than one decision unit, one can base rewards on relative performance; as the number of units increases, under certain circumstances, a first-best optimum can be achieved. (See Barry Nalebuff and Stiglitz, 1983.) We would argue, however, that organizations may perform badly, not only because of misguided intentions, as stressed in the incentive literature, but also from human fallibility. We have been concerned with showing how even in the absence of incentive problems, individual errors are aggregated differently under different organizational forms, leading to systematic differences in organizational performance.

One argument that can be raised against our analysis is that a hierarchy can always decentralize itself, but the converse is not

true; it thus appears, almost tautologically, that hierarchies dominate polyarchies. Within this perspective, the question we have addressed is, under what circumstances should a hierarchy organize itself polyarchically. But we would argue that this perspective is at best misleading: with the right to intervention within a hierarchical structure goes the obligation to intervene when appropriate circumstances arise, and the concomitant necessity to obtain information to effectuate those interventions. Only if there are hard and fast commitments not to intervene, will a hierarchy be equivalent to a polyarchy. The analysis of these issues must, however, await another occasion.

In this paper we couch most of our analysis in terms of a comparison between alternative economic systems; but our results can be applied at a number of different levels of economic analysis (at the level of a firm or an industry as well as for the economy as a whole). Moreover, our results have direct and obvious implications in the context of political decision making, both for the organization of micro decision making (the rules by which committees should operate, or the managerial processes by which public project selection should be conducted) and for the organization of the state. Indeed, we hope our analysis of self-perpetuating organizations, of the problem that all organizations face in selecting those who are to be in decision-making positions, and the comparative sensitivity of organizational performance to the nature of the selection process, will help put into perspective some longstanding fallacies in political theory concerning the virtues and vices of alternative political structures. Classical discussions of the design of State systems (see Karl Popper, 1950) have essentially ignored the problems arising from human fallibility in decision making. Plato, for example, while arguing for the superiority of aristocratic rule, never considered the problems that would arise in choosing the members of the aristocracy over time, or the consequences (by now all too familiar) of the failure to choose well.

In this paper our objective has not been to present definitive results on the comparison

of economic systems. Rather, it has been to encourage a redirection of attention to what seems to us to be one of the most fundamental issues of economics, and to show how simple models can be constructed which provide considerable insights into some of the longstanding controversies concerning the relative merits of different economic systems.

#### REFERENCES

- Lange, Oskar and Taylor, Fred M., *On the Economic Theory of Socialism*, New York: McGraw-Hill, 1964.
- Nalebuff, Barry J. and Stiglitz, Joseph E., "Information, Competition and Markets," *American Economic Review Proceedings*, May 1983, 73, 278-83.
- Plato, *The Republic*; Oxford: Clarendon Press, 1968.
- Popper, Karl, *The Open Society and Its Enemies, Part I—The Spell of Plato; Part II—The High Tide of Prophecy: Hegel, Marx and the Aftermath*, Princeton: Princeton University Press, 1950.
- Sah, Raaj Kumar and Stiglitz, Joseph E., (1984a) "The Architecture of Economic Systems: Hierarchies and Polyarchies," Working Paper No. 1334, National Bureau of Economic Research, 1984.
- \_\_\_\_\_ and \_\_\_\_\_, (1984b) "Economics of Committees," mimeo., Princeton University, 1984.
- \_\_\_\_\_ and \_\_\_\_\_, (1984c) "Perpetuation and Self-Reproduction of Economic Systems: The Selection and Performance of Managers," mimeo., Princeton University, 1984.
- \_\_\_\_\_ and \_\_\_\_\_, (1984d) "Human Fallibility and Economic Organizations: The Architecture of Economic Systems," mimeo., Yale University, 1984.

# Learning from Experience in Organizations

By SCOTT R. HERRIOTT, DANIEL LEVINTHAL, AND JAMES G. MARCH\*

This paper sketches a class of difference equation models for examining incremental experiential learning by economic actors, particularly organizations. The models reflect features of adaptive behavior drawn from observations of decision making in organizations. They picture choice as stemming from decision rules that adjust cumulatively on the basis of trial-by-trial monitoring of the success or failure associated with past adjustments. Such models are in a broad tradition that includes previous work not only in organizational learning and adaptive economics, but also hill-climbing optimization techniques, control theory, and modeling of elementary learning by humans and other animals. Some modest complexities associated with learning are introduced, particularly ways in which learning occurs along several interacting dimensions and within an ecology of learning.

## I. Models of Experiential Learning

We assume a simple choice situation in which a fixed budget is allocated among several alternative, independent activities. Each of the activities provides a return that is proportional to the allocation and the competence (efficiency) of the system at that activity. In the absence of competition, total performance is the potential (or capacity) of each activity weighted by the allocation to that activity and the competence at it, summed over the activities. If  $A_{i,t}$  is the fraction of the budget allocated to activity  $i$  at time  $t$ ,  $k_{i,t}$  ( $0 < k_{i,t} < 1$ ) is the competence

at activity  $i$  at time  $t$ , and  $C_{i,t}$  is the potential return from activity  $i$  at time  $t$ , then  $P_t$ , the performance at time  $t$ , is

$$(1) \quad P_t = \sum_i k_{i,t} A_{i,t} C_{i,t}.$$

Within this choice situation, simple trial-by-trial learning will commonly lead an actor to increase the allocation to activities for which  $k_{i,t} C_{i,t}$  is relatively large, decrease it for those for which it is relatively small.

There are two sets of complications in discovering sensible allocations in this way. The first is that learning occurs along several simultaneous dimensions. Competences and goals adapt at the same time as allocations, and each affects the other. The second complication is that any one learner exists in an ecology of other learners whose actions, goals, and competences affect each other.

## A. Dimensions of Learning

*Adaptive Allocations.* We assume that decision making consists in choosing an allocation ( $A_{1,t}, A_{2,t}, \dots, A_{n,t}$ ) to available alternatives that exhausts the budget. That choice is made by adjusting the previous allocation. The adjustment is made in two steps. At the first step, a proposed allocation to each activity,  $A_{i,t}^*$ , is determined:

$$(2) \quad A_{i,t}^* = A_{i,t-1} + b_1(L_{i,t} - A_{i,t-1}).$$

The learning limit  $L_{i,t}$ , for a proposed allocation to activity  $i$  at time  $t$ , assumes values of 0 or 1 (alternatively 0 and the total budget) with probability  $1 - U_{i,t}$  and  $U_{i,t}$ . Thus,  $U_{i,t}$  is the probability of proposing an increase in the fraction of the budget allocated to alternative  $i$ . The value of  $U_{i,t}$  changes in response to experience, depending on the adjustment in allocation made on the previous trial and the outcome on that trial ( $P_{t-1}$ )

\*Herriott: Department of Management, University of Texas, Austin, TX 78712; Levinthal: Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh, PA 15213; March: Graduate School of Business, Stanford University, Stanford, CA 94305. This research has been supported by grants from the Spencer Foundation, the Mellon Foundation, the Stanford Graduate School of Business, and the Hoover Institution.

relative to some goal ( $G_{t-1}$ ). Specifically,

$$(3) \quad U_{i,t} = U_{i,t-1} + b_2(1 - U_{i,t-1})$$

$$\text{if } A_{i,t-1} > A_{i,t-2}; \quad P_{t-1} > G_{t-1}$$

$$\text{or if } A_{i,t-1} < A_{i,t-2}; \quad P_{t-1} < G_{t-1};$$

$$= U_{i,t-1} - b_2 U_{i,t-1}$$

$$\text{if } A_{i,t-1} > A_{i,t-2}; \quad P_{t-1} < G_{t-1}$$

$$\text{or if } A_{i,t-1} < A_{i,t-2}; \quad P_{t-1} > G_{t-1}.$$

Equation (3) defines a variation on a standard stochastic learning model. Where  $b_2 = 1$ , the direction of the adaptations is determined without stochastic variation. The two learning parameters,  $b_1$  and  $b_2$ , affect the rate at which adjustments in allocations are made. The adjustment of  $A_{i,t-1}^*$  and  $U_{i,t-1}$  from one trial to the next are proportional to the difference between the current values and the upper or lower limits of those variables. They can also be made proportional to the absolute difference between  $P_{t-1}$  and  $G_{t-1}$ , more specifically to  $[|G_{t-1} - P_{t-1}| / (G_{t-1} + P_{t-1})]$ . In such a case, the adaptations are more finely tuned to the magnitude of success or failure. Another form of fine-tuning would disaggregate performance and goal, associating each with a specific activity rather than to their sum. Moving from  $A_{i,t}^*$  to  $A_{i,t}$  involves satisfying the constraint that the sum of allocation increases in individual activities must equal the sum of allocation decreases while maintaining the relative sizes of individual changes projected by  $A_{i,t}^*$ .

*Adaptive Competence.* Models of adaptive learning (for example, binary choice learning, two-armed bandits) commonly assume that the outcome from a current choice is independent of the history of choices. In many situations of economic allocation, however, it seems more reasonable to assume that competence at an activity decreases with the passage of time and increases with allocation to the activity. Thus,

$$(4) \quad k_{i,t} = (1 - b_3)k_{i,t-1}$$

$$+ A_{i,t-1}b_4[1 - (1 - b_3)k_{i,t-1}].$$

The coefficients of competence decay ( $b_3$ ) and learning ( $b_4$ ) control the rate at which efficiency at an activity responds to disuse and experience. Equation (4) is a variant of a standard learning-by-doing model.

*Adaptive Goals.* We assume that performance aspirations adapt to past performance so that the goal in any time period is a mix between the previous goal and the previous performance.

$$(5) \quad G_t = (1 - b_5)G_{t-1} + b_5P_{t-1}.$$

The result is to make the goal an exponentially weighted moving average of performance, with  $b_5$  determining the relative weight attached to relatively recent performance results. If  $b_5 = 1$ , then  $G_t = P_{t-1}$ ; and the adaptation responds simply to changes in performance. If  $b_2 = 1 = b_5$ , the model becomes a standard hill-climbing procedure.

### B. The Ecology of Learning

*Diffusion of Experience.* In a social environment, learning from direct experience is supplemented by the diffusion of experience, that is, by copying others. From a rational perspective, copying can be seen as a way of increasing (on average) the amount of experience from which an individual draws while decreasing (on average) the linkage between that individual's situation and the experience base of action. From a behavioral perspective, it can be seen as a standard way by which adaptive systems deal with uncertainty and ambiguity.

If we let  $A_{j,i,t}^*$  be the proposed allocation of individual  $j$  to activity  $i$  at time  $t$ , a natural extension of (2) yields

$$(6) \quad A_{j,i,t}^* = (1 - d_1)$$

$$\times [A_{j,i,t-1} + b_1(L_{i,t} - A_{j,i,t-1})]$$

$$+ d_1 \sum_{h \neq j} A_{h,i,t-1} / (n - 1).$$

That is, allocations adapt to the mean allocation made by other actors, as well as to direct experience.

If we let  $k_{j,i,t}$  be the competence of individual  $j$  at activity  $i$  at time  $t$ , a natural

extension of (4) yields

$$(7) \quad k_{j,i,t} = (1 - d_2) \{ (1 - b_3) k_{j,i,t-1} + A_{j,i,t-1} b_4 [1 - (1 - b_3) k_{j,i,t-1}] \} + d_2 \max_h k_{h,i,t-1}.$$

That is, competences adapt to the highest level of competence exhibited within the population of actors.

If  $G_{j,t}$  is the goal of individual  $j$  at time  $t$ , then a natural extension of (5) yields

$$(8) \quad G_{j,t} = (1 - b_4) G_{j,t-1} + b_4 \left\{ \left[ d_3 \sum_{h \neq j} P_{h,t-1} / (n - 1) \right] + [(1 - d_3) P_{j,t-1}] \right\}.$$

That is, an actor's goals adapt to the mean performance of other actors, as well as to her own performance. The adaptation coefficients ( $d_1, d_2, d_3$ ) determine the rate at which allocations, competences, and goals spread from one learner to another.

*Interdependence of Experience.* Where experience is interdependent, the performance realized by any one actor depends not only on that actor's allocations and competences, but also on the actions of others. The interdependencies may involve "mating," in which each actor's rewards for a particular activity are augmented by having other actors engaged in the same activity. They may involve competition, in which each actor's rewards for a particular activity are decreased by having other actors engaged in the same activity. They may involve "hunting," in which the rewards of some actors are increased by the presence of other actors engaged in the same activity who, themselves, have their rewards decreased by the joint presence. In the present paper, we consider only the competitive case. If more than one actor allocates effort to a particular activity, the allocation and competence of each reduces the return for the others. Specifically, in any time period where  $\sum_j k_{j,i,t} A_{j,i,t} > 1$ , the return from activity  $i$  for actor  $j$  is

$$(9) \quad R_{j,i,t} = \frac{(k_{j,i,t} A_{j,i,t})^w}{\sum_{h \neq j} (k_{h,i,t} A_{h,i,t})^w} C_{i,t}.$$

The power  $w$  determines the way in which an overexploited activity is shared among competitors.

*Organizational Subunits.* Many economic actors are organizations—firms, armies, public bureaucracies, schools, unions. Organizations have subunits whose actions affect outcomes and whose rewards are linked to local results, as well as to overall performance. We have modeled organizations consisting in subunits, similarly allocating among activities, while adapting allocation, competences, and goals over time. Since allocations within subunits affect not only the performance of subunits but also the performance of the organization as a whole, organizational learning is heavily interactive. The details are omitted here. They parallel the earlier characterizations of learning but include some additional features to link the learning and performance of subunits with the overall organization.

## II. Some Results

We report here some fragmentary results based on analysis of the determinate ( $b_2 = 1$ ) case involving only two alternative activities with unchanging (but different) potentials. We address ourselves to four general questions relevant to assessing trial-by-trial learning as a form of intelligence: 1) To what extent does incremental learning of this type produce, after a suitably long period of time, sensible adaptations to environmental possibilities? 2) To what extent are long-run outcomes independent of initial allocations, competences, and goals? 3) To what extent is the long-run performance of learners improved by increasing the learning parameters? 4) To what extent is the long-run performance of learners improved by tuning the adjustment of allocations more finely to the magnitude of success or failure?

### A. The Isolated Learner

If adjustment of allocations over time is roughly tuned (i.e., if  $b_1$  is fixed), the isolated learner specializes. That is, an equilibrium is reached at which all resources are devoted to one alternative or the other, and where  $k_i =$

$b_4 A_i / (b_3 + b_4 A_i - b_3 b_4 A_i)$ . Specialization is also characteristic of the stochastic version of the model (i.e.,  $b_2 < 1$ ). Specialization may not involve the superior alternative, however. The equilibrium outcome depends not only on the learning parameters but also on the initial allocation, competence, and goal. In general, as the initial conditions become more favorable to the inferior alternative, the set of learning values that result in specialization at the inferior value expands.

Given initial conditions in which competence in and allocation to an inferior alternative are high, specialization in that activity is likely. It can be avoided by slow adjustment of allocations and rapid adjustment of goals, or by learners whose absolute level of performance declines over time (thus producing failures). With fixed capacities for the two alternatives, the latter result requires that the competence decay rate ( $b_3$ ) be high relative to the competence learning rate ( $b_4$ ), that is, slow learning and fast forgetting.

In the "finely tuned" case, where the adjustment of allocations is made proportional to the absolute disparity between performance and goal, the model also reaches a stable mixture of allocations, competences, goal, and performance; but the allocation does not, in general, reach 1 or 0. Rather, it locates an equilibrium combination at some interior, suboptimal point. Thus, fine-tuning yields higher performance in situations in which rough-tuning leads to specialization in the inferior alternative, but not in those cases where rough-tuning leads to specialization in the superior alternative.

### B. Diffusion of Experience

The effects of diffusion of experience among parallel (but noninteracting) learners depends on characteristics of the population of learners. The discussion here is limited to the case of a population heterogeneous with respect to the values of  $b_1$  and  $b_5$ , but homogeneous with respect to the other learning rates (i.e.,  $b_2 = 1$ ,  $b_3 = 0.1$ ,  $b_4 = 0.5$ ). Diffusion of allocations decreases both the mean and the variance of performance within the population (relative to isolated learners), normally driving all actors to a common set

of allocations, competences, and goals. Diffusion of competence, goals, or both normally increases average performance. In addition, the diffusion of competence changes the region of the parameter space that leads to specialization in the superior alternative, giving an advantage in that respect to learners who adjust allocations relatively quickly and adjust goals relatively slowly.

Goal diffusion, by making goals more homogeneous among learners than is performance, tends to divide a population of non-competing learners into one group with a history of subjective successes, another with a history of subjective failures. Since persistent success produces specialization and persistent failure produces nonspecialization, goal diffusion partitions the population into three groups of actors. The first group allocates all its resources to the inferior alternative; the second allocates all its resources to the superior alternative; the allocation by the third group oscillates around an equal division between the two. The proportion in each group depends on the initial conditions and learning parameters. It also depends on whether the allocation adjustments are finely or roughly tuned, with fine-tuning tending to produce a larger number of learners who fail consistently and thus divide their allocations equally among the alternatives.

### C. Interdependence of Experience

The effects of competition depend on the number of competitors and the parameter  $w$  that controls the way in which the resources in an overexploited activity are divided among competitors, as well as the characteristics of the population of competitors. With small numbers of competitors ( $n = 2, n = 3$ ), specialization is common. In the case of two competitors, this means each competitor specializes in a different activity. Under many, but not all, conditions, the slower learner becomes the specialist in the superior alternative, thus has higher performance. In the case of three competitors, a quite typical result is that one or two of the three specialize, while the other does not. Faster learners tend to become specialists, but whether that results in their also having higher perfor-

mance depends on the alternative in which they specialize and the pattern of allocations by the others.

With larger numbers, both the analysis and the story become more complicated. In general, if  $w > 1$ , rapid adjustment of allocations leads to better performance than slow adjustment; if  $w < 1$ , the converse is true. If the adjustment of allocations is made proportional to the magnitude of success or failure (the fine-tuning option), the system reaches an apparent equilibrium which depends on the initial allocations and competences as well as the learning rates. In the fine-tuning case, moderate rates of goal adaptation often seem advantageous, but not always. There are also numerous situations with relatively idiosyncratic outcomes. In otherwise apparently smooth response surfaces mapping variations of performance onto variations in learning rates, substantial spikes appear.

#### D. Competition with Diffusion

If diffusion and competition are both present, we obtain many of the same basic results observed in the case of either alone. Goal diffusion, however, confounds the general observation that rapid, rough-tuned adjustment of allocations gives an advantage where  $w > 1$ . When goals diffuse, competitors who are persistently successful tend to become specialists in one activity or the other. Fast learners tend to specialize, but the fastest learners often specialize in the inferior alternative, leaving the superior alternative to the their somewhat slower cousins. In addition, variation in performance within the population of competitors tends to be decreased by allocation diffusion, but increased by competence or goal diffusion.

#### E. Organizational Subunits

Over a fair range of situations, the consequence of introducing learning subunits is to make both the intelligence of learning through trial-by-trial adaptation and its analysis somewhat problematic. The interactions

make it less likely that organizations will specialize in inferior activities, but also less likely that they will specialize in superior activities. In this respect, the existence of subunits produces effects not unlike the presence of random error in performance. Although it seems likely that there are regular cycles in the resulting patterns of adaptations, we have not as yet discovered them.

### III. Discussion

To provide a base for considering experiential learning as a form of adaptive intelligence, we have modeled a collection of behavioral observations about the forms of learning common in organizations. Since there is ample experimental and observational evidence for believing that simple experiential learning can be a powerful procedure for improving human performance and since the informational, computational, and coordinative requirements of adaptive intelligence seem to be closer to the capabilities of individual and organizational decision makers than are the demands of anticipatory intelligence, we explore the conditions for sensibility of this kind of incremental adaptation.

Although analysis of the models is very incomplete and much of the structure remains unexplored, we can begin to answer the four questions with which the present discussion began. 1) Learning of the sort we have described leads reliably to optimal choices in some situations, but does not do so in others. 2) Allocations at equilibrium are not determined uniquely by activity potentials, but are extensively dependent on the rates at which adaptations take place and on initial allocations and competences. 3) Although fast learners often do better than slow learners, there are many plausible situations in which slow learners do better than fast learners. 4) Although fine-tuned adjustment of allocations facilitates locating an equilibrium, the equilibria achieved are not reliably better than the long-term results of a rougher-tuned adjustment, in fact, are often worse.

# Informational Structure of the Firm

By KENNETH J. ARROW\*

The modern firm is typically not only large but complex. It has an internal structure, and its parts have to communicate and coordinate with each other. This is hardly a new observation. If you will forgive my use of the English platitude, "it's all in Marshall," let me quote, "the development of the organism, whether social or physical, involves an increasing subdivision of functions between its separate parts on the one hand, and on the other a more intimate connection between them" (Marshall, 1948, p. 241).

The complexity of the firm has of course scarcely gone unnoticed in more recent literature. Alfred Chandler (1979) has given a first-rate account of the evolution of the firm's internal structure in response to changing economic needs. Oliver Williamson and others in the bounded rationality tradition stemming from Herbert Simon have sought to create a theory which will accommodate the observed structures of industry. But the history of economic thought suggests that these theories will only find analytic usefulness when they are founded on more directly neoclassical lines, that is, in terms of individual optimization and equilibrium. I do not regard this point of view as some kind of absolute methodological imperative. I merely argue that it has typically been found most convenient to use whether for theory or as a basis for empirical work, and it is worth while to pursue the viewpoint of optimization to see where it will lead.

Nor will I try to give here a full-scale neoclassical theory of internal structure based on specialization and coordination. Rather my aim is to indicate some serious problems in formulating the interchange of information among the component parts of a firm or, indeed, other organization.

I take as the basic model the theory of teams (see Jacob Marschak, 1953, 1954; Marschak and Roy Radner, 1972). Although this theory is now more than thirty years old, its development has been sporadic. It has had as much influence among control theorists as among economists.

The elements of a firm, in team theory, are *agents* among whom both decision making and knowledge are dispersed. The problem, at least as usually formulated, is that of *design*. It is to determine an allocation of information and a set of decision rules for the individual agents so as to optimize some given payoff function for the firm. The payoff depends on the actions taken by all the agents in the firm and on some environmental factors (such as prices or technological conditions) not known at the time the team is designed. These environmental factors are usually referred to as the *state of the world*. The concept of information for an agent is defined as usual in statistical decision theory or communication theory. Each agent observes a random variable, sometimes termed a *signal*. There is a joint distribution of the state of the world and the signals to all agents, which defines a conditional distribution of the state of the world and of other agents' signals given the signal to any one agent.

Each agent has a set of actions from which choice is to be made. Since the agent observes only the appropriate signal, his or her decision rule is a function mapping the agent's signal to that agent's action space. The design problem properly speaking is then both to choose a signal for each agent and a decision rule mapping that signal into actions. We may call the assignment of signals to agents the *information structure* and the choice of decision rules the *decision structure*.

The choice of information structures must be subject to some limits, otherwise, of course, each agent would simply observe the entire state of the world. There are costs of

\*Stanford University, Stanford, CA 94305.



information, and it is an important and incompletely explored part of decision theory in general to formulate reasonable cost functions for information structures. Indeed, most of the theory of teams to date has concentrated on choosing optimal decision structures for a given information structure, rather than optimizing on information structures, thereby avoiding explicit consideration of cost functions.

Up to this point, the model has assumed an initial distribution of information followed by decisions. In the vast literature on economic planning, emphasis is put on communication. Before action is taken, there is an exchange of information, which may take place many times, infinitely often in fact in the analysis of convergent optimization processes. These can be assimilated to team theory. At each stage, the message for each agent to each other agent is a function of the information available to that agent, which now includes the agent's original signal plus all messages received by that agent in previous rounds. What has only now begun to be recognized is that one has to add to an agent's information all the inferences that can be made from the signal and subsequent messages.

The team model does abstract from one very important aspect of organization much stressed in recent literature, that of incentives. It is assumed that the firm as such has organizational objectives which are adhered to by each member. I certainly do not wish to minimize the importance of incentives to perform. There are two reasons why I accept this abstraction here: 1) I want to emphasize the choice of the information structure, which is still of great importance in models with incentives and has been neglected; these models invariably take the structure as given. 2) The present incentive models take a very limited view of the information structure. In fact, within a firm there are many forms of information gathering ("monitoring," in the usual terminology of principal-agent theory) beyond those in our current models. It may be a good abstraction to regard monitoring within the management structure of a firm as sufficiently complete that shirking is not an issue. It is noteworthy that bonuses are to a

large extent discretionary (i.e., tied to observations that cannot easily be quantified) rather than defined functions of measurable performance variables.

With this lengthy background, let me take up some specific problems in characterizing information and communication in firms. First, I will study Martin Weitzmann's (1974) paper on prices vs. quantities as a team problem, to illustrate the analysis in a simple case. I will also use this example for some reflections on the counterintuitive implications of the much-used quadratic payoff functions. Then I want to consider alternative possible assumptions on the cost of information, including the well-known Radner-Stiglitz (1984) theorem suggesting a non-concavity in the production of information. Finally, I will conclude by resuming some recent developments which illustrate some economies in information transfer when the law of large numbers can be used.

Weitzmann proposed the following simple model of decentralization. There are  $n$  productive units, each of which can produce the same product. The cost function of unit  $i$  is

$$C_i(q_i) = t_i q_i + (c/2) q_i^2.$$

Here,  $q_i$  is the output of unit  $i$ ;  $t_i$  is a signal observed by the unit but not by any other agent. Let  $q$  be total output, so that  $q = \sum q_i$ . Let the benefits from a total output of  $q$  be  $B(q) = aq - (b/2)q^2$ .

The random variables  $t_i$  are assumed independently and identically distributed. Without any real loss of generality, it can be assumed that they each have mean zero. Let  $S$  be the common variance of the parameters  $t_i$ . Weitzmann compares two coordination policies. One is to set an output price; each agent maximizes profits given its cost function. The price is chosen to maximize expected benefits minus expected losses. The other policy is that an output quota be set for each firm. Since the prior information is the same for all units, the quota is the same for all units. Unlike the price policy, there is no feedback from the actual signals at all.

Let  $P_p$  and  $P_q$  be the expected payoffs under the price and quantity policies. Then

Weitzmann shows that

$$P_q = na^2/2(nb + c),$$

$$P_p = P_q + [n(c - b)S/2c^2].$$

The team theory solution is to choose the optimal decision rule relating  $q_i$  to  $t_i$ , instead of confining attention to two kinds of rules, that derived from assuming a price and that derived by assuming no dependence of  $q_i$  on  $t_i$ . Because the problem is quadratic, it can be seen that the optimal decision rules are linear. The optimal rule can be found easily (it is in fact a special case of a very general result in Marschak and Radner, pp. 167-69). It can be shown that the expected payoff to the optimal team policy is  $P_t = P_q + nS/[2(b + c)]$ .

As is obvious from its derivation as an optimal policy and as can be calculated directly,  $P_t$  is greater than either  $P_p$  or  $P_q$ . The optimal team policy always lies between the unit outputs derived from the price and quantity policies. Hence, the Weitzmann problem of choosing between the two extremes seems unnecessarily limited.

When examined closely, there are some peculiar aspects to the variation of the optimal payoff with respect to the number of agents,  $n$ , and the variance  $S$  of the cost parameter. The payoff  $P_t$  increases linearly in  $n$  when there is some variance, and so approaches infinity with the number of firms. The same is true of the expected return to the price policy if  $c > b$ . (If  $c < b$ , the expected return to the price policy approaches negative infinity, even more surprisingly.) It is also true that the expected return under the optimal team policy increases with the variance, though possibly this is less paradoxical. The possibilities of gains from trade increase with the diversity of the units.

In fact, all these counterintuitive consequences spring from the quadratic payoff functions together with the failure to observe nonnegativity constraints on the output variables. I take this result to be a warning about trusting quadratic functions too strongly. The quadratic hypothesis has been common in control theory, as well as the economics of

uncertainty, for its analytic convenience. But for some purposes at least it can be very misleading.

This example has taken the information structure as given. If we want to analyze the choice among information structures, it is necessary to include in the analysis measures of their costs. I will give two examples, with applications, before turning to the general question of concavity of the cost function. The first is the Shannon measure of information. Suppose  $X$  is a random variable with a discrete distribution, with  $p(x)$  as the probability that  $X$  takes on the value  $x$ . Suppose we have the following procedure for identifying the true value of  $X$ : at each stage, the remaining possible values of  $X$  can be divided in any desired way into two parts, and it is possible to identify in which part the true value lies. Then a well-known theorem of information theory tells us that the minimum expected number of stages needed to find the true value of  $X$  lies between  $H(X)$  and  $H(X) + 1$ , where  $H(X) = -\sum p(x) \log_2 p(x)$ , the Shannon measure interpreted as a cost. This technology for determining the true value is not entirely convincing, but at least it is a consistent story. As it stands, it is a measure of the cost of going from a probability distribution to certainty. It may be generalized to yield a cost of going from a distribution to a conditional distribution. Let  $S$  be a signal. If  $S$  takes on a particular value, say  $s$ , let  $H(X|s)$  be the Shannon measure for the conditional distribution of  $X$  given  $S = s$ . Then the average reduction in the Shannon measure of uncertainty is,  $H(X) - E_s[H(X|S)] = R(X, S)$ , defined as the *rate of transmission*. This suggests that the cost of a signal  $S$  be taken as proportional to the rate of transmission with respect to the state of the world,  $X$ .

For an application of this measure, I draw upon some earlier work (1971). Suppose an investor can buy a portfolio of *elementary* securities, that is, each security pays off in exactly one state of the world,  $x$ . Let each security have a price of one dollar per unit, and let  $r_x$  be the payment to a unit security for state  $x$ . The investor has  $A$  dollars to invest. The investor's utility function for terminal wealth is taken to be the (natural)

logarithm. The action is the choice of amounts  $a_x$  to be invested in security  $x$ . Terminal wealth then is  $a_x r_x$ . Hence, the chosen portfolio is that which maximizes  $E(\ln a_x r_x)$  subject to the condition,  $\sum a_x = A$ . Straightforward calculation shows that the optimal portfolio is the choice,  $a_x = Ap(x)$ .

Now suppose the investor has the option of buying any signal  $S$  at a price proportional to the rate of transmission, that is,  $cR(X, S)$  for some constant  $c$ . Assume that the signal must be purchased prior to observation of the outcome, so that the purchase price must be subtracted from initial assets before investment. It is easy to see that if a signal with transmission rate  $R(X, S) = R$  is purchased, then the optimal decision rule is to set  $a_x(s) = (A - cR) p(x|s)$ , where  $s$  is the observed value of  $S$  and  $p(x|s)$  the conditional probability that  $X = x$  given that  $S = s$ . The expected payoff with the optimal policy is,

$$E(\ln r_x) + \ln(A - cR) - H(X) + R.$$

Note that this is a concave function of  $R$ . It is maximized with respect to  $R$  by setting  $R = (A - c)/c$  if  $c < A$ ,  $= 0$  otherwise.

Consider an alternative model for suggesting a cost function for information, namely Bayesian normal sampling. The cost of the information obtained from a sample is taken to be proportional to the size of the sample. For a normal distribution, it is reasonable to take the amount of information to be the reciprocal of its variance, called the *precision*. Let  $X$  be normally distributed with precision  $h_x$ . Consider a sequence of signals,  $S_i$ , whose distributions conditional on  $X$  are normal, independent, and identical. Let the conditional precision of any one be  $h_s$ . The agent has to choose an action  $a$ , with loss  $(X - a)$ .

The agent can observe the first  $n$  signals before performing the action. The value of  $n$  is also subject to choice. If  $n$  is chosen, the precision of the conditional distribution of  $X$  given  $S_1, \dots, S_n$  can easily be calculated to be  $h = h_x + nh_s$ . Since the cost of sampling is proportional to  $n$ , it can also be seen to be linear in  $h$ . If we assume that the cost of sampling is added to the loss due to the

action, it can be seen that the agent wants to choose  $h$  to minimize  $h^{-1} + ch$ , again a concave function. The optimal value of  $h$  is  $c^{-1/2}$ .

Radner and Stiglitz have proved the following remarkable theorem:

Let  $S(t)$  be a one-dimensional family of signals defined for  $t \geq 0$ . The conditional density of  $S(t)$  given  $X$  is  $p_t(s|x)$ , assumed differentiable in  $t$  at  $t = 0$ . Further assume that  $p_0(s|x)$  is independent of  $x$ , that is,  $S(0)$  yields no information about  $X$ . Let  $a(s, t)$  be the action taken if the signal  $S(t)$  has been chosen and takes on the value  $s$ . Let  $w(a, x)$  be the outcome in money terms if action  $a$  is taken, and the state of the world is  $x$ , and let  $c(t)$  be the cost of having the signal  $S(t)$ ; assume that  $c'(0) > 0$ . Finally, let  $V(w - c)$  be the utility function for income. Then the expected utility gain by changing the signal from  $S(0)$  to  $S(t)$  is always negative for  $t$  sufficiently small.

Alexander Pope told us, "a little learning is a dangerous thing." Radner and Stiglitz, a little more circumspectly, tell us that a little information is never worth the cost. This conclusion implies that the value of information is not concave, at least not in the neighborhood of the origin, which corresponds to an uninformative signal. But in our two examples, both certainly reasonable, the value of information was concave in a properly chosen measure. The Radner-Stiglitz theorem is correct mathematically, so that we can be sure that one of its hypotheses does not hold in the two examples. The hypothesis that fails is the seemingly innocuous condition that the likelihood function  $p_t(s|x)$  is differentiable in  $t$  at  $t = 0$ . If we parametrize the signals in our two examples by  $R$  or  $h$ , respectively, then  $p_t(s|x)$  has an infinite derivative with respect to  $t$  at  $t = 0$ .

I conclude by discussing a different kind of subtlety in the use of information. Team theory differs from the classical work on the economics of socialism in several ways, but one is the use of a priori information in guiding the communication and coordination processes. One particular kind of argument that has been developed is to take advantage of the law of large numbers when the agents form a well-defined statistical universe. In a

resource allocation problem, when the center's action is the allocation of some scarce resource among units, the optimal action will in general depend on the productivities of all the units. However, if these productivities can themselves be considered as a random sample from a known distribution and if there are many units, then effectively the distribution in the sample can be regarded as known. Therefore, some information need not be collected, or, more precisely, its value goes to zero as the number of agents approaches infinity. A model of this kind has been studied by myself and Radner (1979), and similar reasoning has been used independently by J. K. Leenstra et al. (forthcoming), though more for the purposes of approximating the solution of nonconcave problems.

These remarks are just a few of the many possibilities which can be invoked in theory to analyze the optimal informational structure of firms. They suggest that the solutions in different circumstances may look very different, even though based on the same maximizing principles.

#### REFERENCES

- Arrow, Kenneth J., "The Value of and Demand for Information," in C. B. McGuire and R. Radner, eds., *Decision and Organization*, Amsterdam: North-Holland, 1971, 131-39.
- and Radner, Roy, "Allocation of Resources in Large Teams," *Econometrica*, March 1979, 47, 361-85.
- Chandler, Alfred, Jr., *The Visible Hand*, Cambridge: Belknap Press, 1979.
- Leenstra et al., J. K., "A Framework for the Probabilistic Analysis of Hierarchical Planning Systems," *Annals of Operations Research*, forthcoming.
- Marschak, Jacob, "Équipes et Organisations en Régime d'Incertitude," in Centre National de la Recherche Scientifique, *Économetrie*, Paris, 1953, 202-11.
- , "Towards an Economic Theory of Organization and Information," in R. M. Thrall et al., eds., *Decision Theory*, New York: Wiley & Sons, 1954.
- and Radner, Roy, *Economic Theory of Teams*, New Haven: Yale University Press, 1972.
- Marshall, Alfred, *Principles of Economics*, 8th ed., New York: Macmillan, 1948.
- Radner, Roy, and Joseph Stiglitz, "Nonconcavity in the Value of Information," in M. Boyer and R. E. Kihlstrom, eds., *Bayesian Models in Economic Theory*, Amsterdam: Elsevier, 1984.
- Weitzman, Martin L., "Prices vs. Quantities," *Review of Economic Studies*, October 1974, 41, 477-92.

## INDUSTRIAL POLICY IN FRANCE<sup>†</sup>

### State and Industry in France, 1750–1914

By CAGLAR KEYDER\*

For most students of the subject, French economic history is, as Shepard Clough called it, the “history of national economics” (1939). Especially in the Anglo-Saxon tradition, it is the otherness, or the contraposition to the presumed normalcy of the English case which draws attention to the history of the French economy. It is, of course, not accidental that the Anglo-Saxon tradition has upheld the British experience as the ratio since the successive hegemonic powers (Britain and the United States) defended, at least in their ideological posture, what is supposed to constitute the sufficient institutional basis of successful industrialization, viz, the autonomous market.

The French experience, on the other hand, despite its providing a model with a much more justifiable claim to universality, has been treated mainly in its distinction from the market paradigm of the first industrial country. Presumably it was the reversal in economic supremacy which became all too obvious in the 1970's that prompted revisionist work reassessing French industrialization. The principal contributions are by Richard Roehl (1976), and Patrick O'Brien and myself (1978). For reviews, see Rondo Cameron and Charles Freedman (1983), and N.F.R. Crafts (1984). There is, as yet, no attempt on the part of economists to draw the conclusion from this revisionism to advocate industrial policy and a state-economy relationship derived from the French experience.

What is of concern is the role of the state in the economy, or euphemistically, the nature of economic policy. At some essential level, France is the privileged locale where the state is visible and, to varying degrees, autonomous. This is not only a retrospective judgement: it was a constant theme in contemporary discussions as well, which aimed at discovering the best role for the state in constructing the desired national economy. At this level, fluctuations in French economic policy may be presented as various expressions of the essential nature of the state and its relation to the economy. From Nef through Boissonade, Cole, Heckscher, and Clough, the theme that runs through the accounts is *étatisme* whether in mercantilist, Colbertist, or interventionist and protectionist garb. There are certainly excesses and departures from the general line, but the continuity of the importance of the state is what sets France apart.

#### I

There are three not entirely independent explanations for the persistent importance of the state in French economy. The first focuses on the role and demands of the economically dominant classes—merchants, shippers, and industrialists. Confronted with British superiority, these groups appealed to the state for protection and the state responded with economic policy designed to secure the internal market and gain a better share in the world market for its prime constituency. In other words, French specificity results from the British economic threat and the domination of the French bourgeoisie over, or their collusion with, the political authority.

The evidence of a secular French inferiority in industry is far from conclusive, except in sectors which constituted the mainstay of

<sup>†</sup>*Discussants:* W. James Adams, University of Michigan; John Zysman, University of California-Berkeley.

\*Department of Sociology, State University of New York, Binghamton, NY 13901.

the industrial revolution, but French industrialists' unreasonable fear of British competition is well documented. On the other hand, it was not only industrialists who appealed to the state for protection against British competition, but also various colors of social reformers who were concerned that France should not repeat the British example of a socially disruptive, costly, and morally repugnant industrialization. In both cases, British presence was utilized to remind the political authority of its duties toward society and the economy. The main problem with this explanation is its instrumentalization of the state in the hands of the economically dominant groups. Cases such as the 1786 and 1860 treaties raise a difficulty with this view concerning the explanation of policy acts which ostensibly opposed bourgeois interests. This problem may be solved, however (though, I fear, unsatisfactorily), through a better understanding of the fractions within the bourgeoisie and their relative economic standing. The argument would then become that, with the exception of a few instances, dominant factions in economically dominant classes favored an extensive presence of the state in the economy, primarily for the sake of strengthening against and protecting from British industry.

The second perspective which seeks to define an essential pattern of the state-economy relationship extends the tutelage of the bourgeoisie under the political authority to a permanent dimension of the French social formation. In this view, it is the genetic strength of the French state, as it was constituted during the crisis of feudalism, and as it later perpetuated itself, which determines its relations to the economy.

The relative weakness of the landlords and the concomitant strength of the late feudal monarchy made for the consolidation of peasant property, ratified in the Revolution. By contrast, absolutist centralization failed in England where the aristocracy expropriated the peasantry and dominated the state. The syncretic relationship between the peasantry and the state allowed for the French monarchy and the bureaucracy first to withstand enclosure tendencies by the aristocracy and later, as exemplified in the Sec-

ond Empire, to establish an extensive state machinery relatively autonomous from the bourgeoisie. The central authority implemented its external and internal goals, such as military strength and domestic stability, necessarily through the manipulation of the economy. In other words, it has been a permanent feature of the French state to construct and guide an economy in its design. Bertrand Badie and Pierre Birnbaum (1983) provide the most explicit version of this "autonomy of the state" approach.

An objection to such an autonomization of the state would be expected from the orthodoxy seeing in the state a reflection of changing social balances, and more specifically from the proponents of the classical interpretation of the Revolution: for, if the bourgeoisie did not capture the state, then what exactly did the Revolution accomplish? If the state-economy relationship is unchanging, is there then no substantive difference between the bourgeoisie's relationship to state power in the eighteenth and nineteenth centuries? As is well known there is an active current, ranging from Furet to Wajda, attempting to de-construct the Revolution, and an argument for the constancy of the nature of the state would necessarily be revisionist in a similar vein. For a review of this literature, see Jack Goldstone (1984).

This revisionism aims at debunking what Furet called the "revolutionary catechism," but in doing so it tends to ignore the dimension of eighteenth-century reality whose transformation permitted the postrevolutionary state to engage in a qualitatively different economic policy. The eighteenth century had witnessed a dislocation of the privileged relation between the peasantry and the state to allow for tax-farmers and rentiers to indirectly exploit the peasantry. The state had been reduced to a role of intermediation between producers and a money bourgeoisie, organized at a central level. On the eve of the Revolution, three-fifths of the state budget served to pay the interest on public debt, thus curtailing the political authority's ability to implement any substantive transformatory policy at the level of the economy. This problem, which became manifest at the level of state finances through a fiscal crisis, is

widely considered to be one of the proximate causes of the Revolution, but it is not stressed that the Revolution dealt a decisive blow to this class of rentiers through the inflation of the *assignats*. Thus, in addition to laying out the institutional structure and the juridical forms of capitalism, the Revolution permitted the state to engage in economic policy to cultivate the bourgeoisie. This perspective suggests that the eighteenth-century state was heavily mortgaged to a stratum which could not be construed as the subject of an economic policy oriented to the construction of a national economy, and with the repudiation of the *assignats* in 1797, the state was freed of this indenture to return to its Colbertist nature. In other words, the Revolution may be interpreted as restoring the state to its essential status in the society.

Colbert has been frequently invoked as having "established a strong and virtually indestructible tradition in French economic policy which the bourgeoisie was to use and interpret in its own interests" (Tom Kemp, 1971, p. 34). In the sense of a tradition lending legitimacy to various measures of industrial policy, Colbertism was as good a rallying cry as any; however, it took upon a social dimension as well.

During the nineteenth century, the argument was frequently advanced by protectionists that industrial policy was the principal guarantor of stable employment and what they considered high wages in the manufacturing sector. A drift from the relative insularity of industry would necessitate rapid change leading to unemployment and lower wages and threaten the social compact. This might seem to be a strange argument since revolutions occurred nonetheless and since average wages were already appreciably lower than their British counterparts. (All the quantitative comparisons with Britain derive from the book by O'Brien and myself.) Yet, there might be some justification to the claim from the perspective of rural-urban income differentials. With a low demographic growth and a secure property regime based on peasant ownership, the push factor does not operate intensely, thus average incomes in agriculture plus the risk involved in abandoning them, have to be matched in the city for there to be any migration. Protection and

industrial subsidies probably allowed for maintaining higher wages than would have been the case under free trade. More importantly, however, protection permitted the survival of an industrial structure based to a larger degree than Britain on artisanal and small producing units. This combination of Colbertian tradition and a coalition formed by a variety of interests including the grande bourgeoisie, small capital, artisans, and workers induced the state toward conservatism when it came to the policy of protection. (See, for example, the essay by Claude Fohlen, 1956.)

Liberalization was widely considered an English plot locally supported only by Girondin merchants. By preventing social and economic dislocation, the state defended the status quo, and consequently industrial policy was characterized by a remarkable continuity of purpose. According to this third explanation, it was the historical formation of the society as a whole which induced its collective expression—the state—to protect it from disruptive forces. Since such forces were widely recognized as economic and external, the state had to take upon a prominent policymaking role in its relationship to the society. If the first explanatory attempt was orthodox and class reductionist, the second may be labelled politician and state-autonomist. The last one, by contrast, is pluralist as it appeals to the conjunction of all social interests.

## II

A second type of approach eschews characterization of the permanent nature of the state-economy relationship and instead engages in identifying successive periods punctuated by political events. The politicized character of the French state, especially in the nineteenth century, is reminiscent of Third World experience in the twentieth. Régimes were temporary; political institutions did not enjoy effective legitimation; and the system was questioned at every crisis. Policy change correlated with regime change.

The Revolution was the first fundamental break; soon followed the inception of the continental system and Waterloo, then the

coup d'état of 1830, the 1848 Revolution, the 18th Brumaire of Louis Bonaparte, the Commune and the German War, and the Third Republic. For each of the periods in between, competing interpretations of the state-economy relation are available and they usually refer to the government being more or less responsive to the bourgeoisie, or to particular bourgeois factions as opposed to others.

The periodization actually begins with the ancient regime; the state in the period after 1750 came under Physiocratic influence and was more recognizant of manufacturers' needs and, despite the adherence to the old instruments, "the intendants tended to become more concerned with encouraging entrepreneurs than with enforcing regulations" (Kemp, p. 72). Kemp's judgment may be opposed by this conclusion of Pierre Deyon and Philippe Guignet: "There is no reason for believing that manufacturing policy was radically transformed between the first and second halves of the century" (1980, p. 629).

The 1786 trade treaty with England was definitely an instance of liberalization although its application period was very short. Even if it was not the immediate cause of the pre-Revolutionary crisis, this treaty provided the Republicans with an ideological banner. The bourgeoisie wanted not only the constitution of a liberal economy based on national integration and private property, but also protection against foreign industrial and commercial competition. What was demanded in the *Cahiers* was rapidly fulfilled in the unfolding of the Revolution. First, of course, came the sanctity of private property among other juridical constellations and measures toward the economic integration of the country.

In 1792, the war and the defeat of the Girondin faction allowed the Convention to abrogate the Eden treaty and, shortly thereafter, to announce the start of economic warfare against England, which prohibited all British goods from entering French soil. The true promotion of the manufacturing bourgeoisie came through the economy of war-time mobilization and the Blockade. In the attempt to construct an autarkic Continental System, many industries which would have been doomed to failure under any

world-economic sanctioning prospered and must have allowed their entrepreneurs opportunities of rapid accumulation. The military end to the Continental System coincided with its internal dissolution through shortages, smuggling, and trade licenses granted by the government: the bourgeoisie had begun to find Napoleon's political direction of the economy dysfunctional. (On this subject, Eli Heckscher's 1922 book remains a valuable summary.)

Economic policy under the Restoration may be seen as a return to less radical protectionism, without the supplanting of the market that the Napoleonic system had resorted to. The new tariffs reflected the composition of the Chamber, where returning landowners and the bourgeoisie sat; once again the exporting interests of Bordeaux had been excluded from the coalition. They remained so under the July monarchy as well, which has been called the period in France when the state was the least autonomous from the grande bourgeoisie (Jean Lhomme, 1960, p. 57).

The grande bourgeoisie was a newly ascending class yet it successfully opposed the Bourbons who represented to a public, which now remembered only the glory of Bonaparte, the defeated Ancien Régime. Under Louis Philippe, bankers held actual power: as Marx described it, the July monarchy was a business concern for the finance aristocracy. They plundered the state budget through floating loans and railroad schemes; they prevented financial reform and rational bourgeois calculation. Nevertheless, the bankers succeeded in increasing credit and there was considerable industrial expansion even if not at revolutionary pace. The success of the financiers swelled the ranks of the bourgeoisie excluded from governmental posts and the numbers of industrial workers who joined forces in the 1848 revolution to overthrow the monarchy. The Second Republic accordingly aimed at neutralizing the power of finance by strengthening the Bank of France.

The subsequent phase could have been more directly dominated by the industrial bourgeoisie as the revolution seemed finally to have run through its arduous course: an attempt by Napoleon to build an economy



under political auspices, a short-lived restoration, a usurpation by the grand bourgeoisie of the liberal precepts punctuated by 1848. The factor of Bonapartism intervened.

The regime of Louis Bonaparte is a favorite of social scientists where a name provided an otherwise paltry character with charisma and the "normal" political representation of the economically dominant classes was interrupted. Relying on a mass constituency of peasants, workers, and the lumpen of Paris, and ruling by plebiscite, the authoritarian Bonapartist state could afford to be autonomous of bourgeois interests; the bureaucracy and the state machinery which had been centralized and rationalized under the uncle, attained their own power under the nephew.<sup>1</sup>

For Karl Marx, Bonapartism represented another effect of the overwhelming weight of the small peasantry as it determined the state-society equation (1963, pp. 128-31). By responding to the ideological aspirations of the small peasantry, Louis Napoleon could break the political power of the bourgeoisie. Under him, the administrative machinery increased in size, the bureaucracy became powerful, institutionalized and differentiated, and evolved into a self-sustaining cadre with its own schools and recruitment practice. The autonomization of the bureaucracy allowed for policies which were distinguished by their bolder approach to economic modernization. No longer ridden with conservative solidaristic fears of an industrial future, the government engaged in long-term financial policy, supported railway construction and rebuilt Paris under the guiding influence of Saint-Simonian bankers. Lifting official restrictions on the formation of joint-stock companies, an impetus was given to industrial investment and the transformation of industrial structure toward concentration. All such measures may be interpreted within the framework of a state structurally bound to capitalism, but enjoying an instrumental

autonomy from particular groups within the bourgeoisie.

In light of what seemed to be a unanimous opposition to trade liberalization, the 1860 Cobden-Chevalier treaty with the principal economic rival across the Channel and subsequent treaties with other European countries is the most surprising act in the Bonapartist repertoire. This was truly a revolution from above where, according to one interpretation, Napoleon was convinced that he could appraise long-term interests of the bourgeoisie, and despite complaints in the short run, the bourgeoisie, with the exception of a few intransigent protectionists, were grateful. (Marcel Rist, 1956, and Lhomme both endorse this interpretation.) Yet by the end of the 1860's, the general will had begun to crumble and the Third Republic witnessed a return to many of the older constants of economic policy.

The relative political stability of France until World War I represents a new equilibrium between the state and the economy. During this period, neither the bourgeoisie nor the bureaucracy were unilaterally capable of formulating policy. Nationalist sentiment after the German war had created a consensus behind a gradual departure from free trade, and especially following the crisis ushered in by the 1873 financial collapse, agricultural and industrial interests could meet on a common ground albeit under the latter's domination. Soon, however, the agricultural depression and its more general consequences in the economy created sufficient pressure in the direction of conserving France's agrarian structure. The peasantry was once again saved through the Méline tariff, and once again, as Eugene Golob says, "the fate that had befallen English agriculture" (1944, p. 9) was avoided. A sturdy peasantry would ensure that industry and agriculture could enjoy a balanced growth, in conformity with the national economic teaching of the day.

The economic boom that followed the 1892 tariff seemed to confirm protectionist hopes, but it also brought with it an increased willingness by various groups within the bourgeoisie to entrench their particular interests into state policy. The parliamentary system,

<sup>1</sup>The concept of the Bonapartist state derives from Marx's work (1963); Nicos Poulantzas (1973) gave it new currency in the late 1960s. For a review of the literature in France, see Alain Rouquié (1975).

by now stabilized, provided the political structure for such contestation and closure. As the French economy lost its relative share in world trade, partly due to the tariff, the boom built on her peculiar social structure continued and it was precisely in the two decades preceding World War I that total commodity output increased to exceed Britain's. Industrial output as a proportion of Britain's had declined from 1.3 around 1860 to 0.65 by late 1890's; the 1905-14 average increased to 0.7.

### III

I have, thus far, examined the state-economy relationship as manifested in sharp breaks in general policymaking. There is, however, another realm where the shifts are not so visible, but the political input into the economy is at least as important—that of state expenditures. From the above discussion it would be expected that the periodization based on regime changes with alternating perspectives on economic policy would also be a good predictor of state expenditures, where, for instance, increasing fiscal presence correlated with attempts at building the national economy. It appears, however, that state expenditures vary according to a different periodicity, one taking shape largely outside the French economy.

Louis Fontvieille (1976), in his construction of statistics on state expenditures in the ISEA series which include Toutain's estimates on agriculture, Markovitch's on industry and Marczewski's on physical output, finds that state expenditures as a proportion of physical product, show a secular tendency to increase, from around 10 percent until 1870 to around 15 percent before World War I. More importantly, however, the share in state expenditures of what might be called social investment, or what Fontvieille terms "expenditures related to the development of the national economy" increase from 3-4 percent in the first half of the century to 7-8 percent. This evolution is as expected; what is surprising in the fluctuations around the long-term trend is the author's finding that Kondratieff cycles are correlated with the stagnation or the spurts in the ratio. Thus,

during *A* phases of economic prosperity (1848-73 and 1896-1914) the ratio is stable or even declines, while in *B* phases of depression (1815-48 and 1873-96), it increases rapidly.<sup>2</sup> In other words, successive governments, regardless of the character of the regime, seem to have practiced countercyclical policy, in particular an injection of state expenditure into the industrial social capital. For Fontvieille, this finding allows the conclusion that the development of the state apparatus is determined by the devaluation and accumulation needs of capital in the successive stages of development of industry.

This explanation of the movement of state expenditure has been challenged and debated especially on grounds that the fluctuations in French economy do not exactly correspond to standard Kondratieff cycles. Although it does not constitute sufficient proof of the state being an instrument of the bourgeoisie, the correlation Fontvieille has discovered remains strong. In one case, that of the Plan Freycinet, a public works program which was initiated in 1878, the intention of counteracting the crisis of industry, especially in the iron and steel sector, was clear.<sup>3</sup> Although the Plan Freycinet's extraordinary budgets ended in 1882 and financial orthodoxy reasserted itself, state expenditures as a proportion of physical output continued to grow until the end of the Great Depression.

### IV

Long-term structural concerns emphasizing permanence and attempts at accounting

<sup>2</sup>"Related expenditures" consist of public works, social and economic activities, education, colonial expenditures and interest payments on the public debt. According to Fontvieille, the ratio of related expenditures to physical output increased from 2.4 to 4.4 percent between 1815 and 1849; from 5.4 to 8.9 percent between 1870 and 1895; while between 1850 and 1869, it only moved from 3.9 to 4.1 percent, and from 1895 to 1913, it actually declined from 8.9 to 7.1 percent (p. 1673).

<sup>3</sup>Yasuo Gonjo's article (1972) on the Plan Freycinet makes its industrial market creation dimension very clear.

for observed fluctuations through a periodization need not be alternatives. They bear the same complementary relation to each other as analyses of *longue durée* and *conjecture*. There was, in fact, a distinctive social structure in France during the development of its capitalism. This structure determined a certain institutionalization and autonomy of the French state which shaped its absolute presence and the instruments it possessed in intervening in the economy. And this state form evolved, after Colbert, in a direction which established the bureaucracy's relative strength and ability to implement their visions of the national economy. Regimes, on the other hand, changed rather rapidly, without, however, altering the basic state-economy relationship. They did allow more or less representation to the bourgeoisie, arguably in relation to world and national economic conjuncture; and it seems that regime changes sometimes led to a ready assumption of policy change.

Specific policies may not be seen as simply deriving from either the essential nature of the state or the character of the regime: it was the interaction between the momentum of the state originating indeed in the absolutist consolidation (Colbertian étatism and its mercantilist assumptions of rivalry) and the conjunctural needs of industrial capitalism (as these were reflected in political regimes) which shaped them.

#### REFERENCES

- Badie, Bertrand and Birnbaum, Pierre, *Sociology of the State*, Chicago: University of Chicago Press, 1983.
- Cameron, Rondo and Freedeman, Charles E., "French Economic Growth: A Radical Revision," *Social Science History*, Winter 1983, 7, 3-29.
- Clough, Shepard B., *France: A History of National Economics, 1789-1939*, New York: Charles Scribner's Sons, 1939.
- Crafts, N. F. R., "Economic Growth in France and Britain, 1830-1910: A Review of the Evidence," *Journal of Economic History*, March 1984, 44, 49-67.
- Deyon, Pierre and Guignet, Philippe, "The Royal Manufactures and Economic and Technological Progress in France before the Industrial Revolution," *Journal of European Economic History*, Winter 1980, 9, 611-32.
- Fohlen, Claude, "Bourgeoisie Française, Liberté Économique, et l'Intervention de l'État," *Revue Economique*, May 1956, 3, 414-28.
- Fontvieille, Louis, "Evolution et Croissance de l'État Français de 1815 à 1969," published as a separate issue of *Economies et Sociétés*, September-December 1976, 10.
- Goldstone, Jack A., "Re-interpreting the French Revolution," *Theory and Society*, September 1984, 13, 697-713.
- Golob, Eugene O., *The Méline Tariff: French Agriculture and Nationalist Economic Policy*, New York: Columbia University Press, 1944.
- Gonjo, Yasuo, "Le 'Plan Freycinet', 1878-1882: Un Aspect de la 'Grande Dépression' Économique en France," *Revue Historique*, 1972, 1, 49-86.
- Heckscher, Eli F., *The Continental System; an Economic Interpretation*, Oxford: The Clarendon Press, 1922.
- Kemp, Tom, *Economic Forces in French History*, London: Dennis Dobson, 1971.
- Lhomme, Jean, *La Grande Bourgeoisie au Pouvoir (1830-1880)*, Paris: Presse Universitaire de France, 1960.
- Marx, Karl, *The Eighteenth Brumaire of Louis Bonaparte*, New York: International Publishers, 1963.
- , *Class Struggles in France, 1848-1850*, New York, 1964.
- O'Brien, Patrick and Keyder, Caglar, *Economic Growth in Britain and France, 1780-1914*, London: Allen and Unwin, 1978.
- Poulantzas, Nicos, *Political Power and Social Classes*, London: New Left Books, 1973.
- Rist, Marcel, "Une Expérience Française de Libération des Échanges au Dix-Neuvième Siècle, le Traité, de 1860," *Revue d'Economie Politique*, November-December 1956, 66, 908-61.
- Roehl, Richard, "French Industrialization: A Reconsideration," *Explorations in Economic History*, July 1976, 13, 233-81.
- Rouquié, Alain, "L'Hypothèse 'Bonapartiste' et l'Émergence des Systèmes Semi-Compétitifs," *Revue Française de Science Politique*, December 1975, 25, 1077-111.

# French Industrial Policy under the Socialist Government

By BELA BALASSA\*

In presenting the draft law on nationalizations in French industry and banking to the National Assembly on September 27, 1981, "the lack of a true industrial policy" was adduced as the principal reason for the proposed actions. It was claimed, in particular, that "it is necessary for the state to have the instruments for efficient interventions and for the planned orientation of the country's development. The most important of these instruments is the enlargement of the public sector." The enlarged public sector was to be the "fer de lance" of modernization of the French economy, with the state providing the funds for the requisite investments. Modernization appeared as the key word in statements made by socialist leaders, who repeatedly claimed that "there are no condemned sectors, only outdated technologies."

The objective of developing simultaneously all sectors while improving technology was reiterated in a speech before the chief executives of public enterprises in the industrial sector on August 31, 1982, by Jean-Pierre Chevènement, then Minister of Research and Industry. A year later his successor, Laurent Fabius, now Prime Minister, suggested in a speech made to the National Assembly on October 31, 1983, that "an industrial strategy involves the choice of the principal national priorities...in pursuing two major objectifs: to contribute to the re-establishment of economic equilibria in particular in employment, and to modernize the industrial structure."

This paper will review the French record with industrial policy since May 1981. It will

examine the experience of the major industrial sectors, discuss the actions taken, evaluate the results of these actions, and indicate possible future prospects.

## I. The Experience of the Major Industrial Sectors

Coal, steel, and shipbuilding, three declining industries, exhibit a similar pattern. After May 1981, a reversal occurred in the process of restructuration that would have entailed continued reductions in output and employment. The targets were raised substantially above actual production levels, leading to new hiring. With declining demand and the rise of wages and social changes, including a fifth week of vacation and full compensation paid for the reduction of the work week from 40 to 39 hours, there resulted a considerable increase in the deficits of the firms in the three industries.

Notwithstanding the rise in budgetary appropriations, the firms in question had to increase their borrowing, thereby raising levels of indebtedness. By 1984, long-term debt surpassed the value of output in the coal mining industry, it was one-half of output value in steel making, and in between these two figures in shipbuilding. In the same year, deficit and budgetary support combined amounted to 140 thousand francs per worker in coal, 115 thousand francs per worker in steel, and 175–200 thousand francs per worker in shipbuilding. By comparison, the average wage in the manufacturing sector was 110 thousand francs.

High financial charges on their debt will continue to add to the losses of the firms in the three industries. Further costs are involved in connection with the so-called congés de conversion, under which the workers who are considered superfluous are receiving over a two-year period, 70 percent of their pre-tax salary, financed one-third each by the firms themselves, from the government budget, and by the unemploy-

\*Professor of Political Economy, Johns Hopkins University, Baltimore, MD 21218, and Consultant to the World Bank. I am indebted to French economists, businessmen, and government officials for helpful discussions on the subject matter. I alone am responsible, however, for the opinions expressed herein; they should not be interpreted to reflect the views of the World Bank.

ment fund. Also, in some cases, substantial bonuses are offered to workers who depart voluntarily (50,000 francs in one of the large shipyards).

The automobile industry was traditionally an important export sector in France. The position of the industry deteriorated to a considerable extent in external as well as internal markets after May 1981, however. Thus, its share in European markets fell from 30.4 percent in 1980 to 24.3 percent in 1983, whereas the share of foreign cars in France rose from 21.7 to 32.7 percent. During the same period, the combined losses of the Peugeot and Renault groups increased from 1.2 milliard to 4.3 milliard francs.

Several factors contributed to these results. According to estimates cited in the report of the Conseil Économique et Social, the measures taken after May 1981 increased labor costs in the automobile industry by 19.5 percent in 1982. At the same time, the price control introduced in June 1982 did not allow for offsetting increases in prices, giving rise to foregone revenue of about 2.5 percent. Finally, social troubles, fostered largely by the C.G.T., involving strikes, production slowdowns, and the deterioration of product quality, contributed to declines in market shares and added to losses, which were increased further as the labor unions obstructed reductions in the industry's work force.

The losses incurred have in turn given rise to high financial charges, accounting for about 4 percent of total costs in the two groups. At the same time, their long-term debt came to exceed two-fifths of the total value of sales, with the ratio of the debt to capital approaching 2, compared to ratios between 0.2 and 0.9 for major competitors abroad.

The deterioration of the financial position of the automobile companies has not permitted undertaking the necessary investments in automation that have been carried out by foreign competitors. The difficulties encountered in reducing employment have also discouraged automation and labor productivity practically stagnated between 1980 and 1983 in the French industry, compared with gains of about 20 percent in General

Motors, Ford, and Fiat. These results, cited in a report, prepared in the summer of 1984 by a commission chaired by Francois Dalle, the president of the L'Oréal cosmetics group, reflect excessive manning levels in French automobile factories. According to the same report, the work force of 230 thousand in the industry in June 1984 should be reduced to 160 thousand by 1988.

Among high-technology industries, France made considerable progress in nuclear energy and in aerospace during the 1970's. Its record in electronics was mixed, with strengths in telecommunications and in professional electronic equipment, and weaknesses in electronic components, such as integrated circuits, in computers, and in electronic consumer goods. On coming to power, the socialist government gave considerable emphasis to electronics, with a view in particular to strengthen the weak areas. It further set out to develop fiber optics for use in a nationwide network of cable transmission.

The plan "filière électronique" envisaged spending 140 billion francs (in 1982 prices) on investment and R&D in electronics over a five-year period. Within this total, 11-12 billion francs a year were to have come from the government budget. In fact, 9.5 billion francs were spent in 1983 and an estimated 11 billion francs in 1984, but in current rather than in constant prices. Although comparable data are not available, it appears that the shortfall has been even greater as far as expenditures from nongovernmental sources are concerned.

The spending actually undertaken nevertheless represents an increase over the 1970's, permitting the French electronics industry to reach a rate of growth of output of 8 percent in 1983. However, in the same year, the rate of expansion of output was 16 percent in Japan. Furthermore, several questions arise in regard to the future prospects of the French electronics industry.

To begin with, one may query the desirability to "faire cavalier seul" in integrated circuits and computers. A similar effort failed earlier in the so-called Plan Calcul. In both industries, technological change is very rapid and alliances with foreign firms provide advantages for the development of new tech-

nology; European firms outside France have in fact entered into such alliances with American and Japanese firms.

Further questions arise concerning the planned expansion of the production of electronic consumer goods in France. Alongside the failures, such as the Plan Calcul and Concorde, French industrial planners were successful in the past in developing branches of electronics where both supply and demand were dominated by the state, thereby imparting considerable stability to the market.

In the case of consumer electronics, however, tastes change rapidly and centrally made plans easily go awry. Also, government interventions in production decisions, such as compelling Thompson to build a factory producing parts for video recorders in Longwy, interfere with firm decision making and increase costs. And, while in France the restructuring of the nationalized firms has led to the establishment of monopoly positions, the American experience indicates that small firms have the advantage of flexibility in catering to consumer needs and competition becomes a force of progress.

France has also chosen a highly centralized alternative for establishing a nationwide cable network, to be controlled and financed by the Direction Générale des Télécommunications of the PTT. By contrast, in Britain, there will be competing cable networks to be financed privately, as in the United States.

With the deterioration of the financial position of the PTT and the policy of austerity, only about 2 billion francs of the total of 60 billion have been spent so far on experimental schemes in a few cities in France, and considerable delays have been experienced compared to the original plan. At the same time, the installation of the fourth TV channel (Canal-Plus) on a subscription basis may have preempted the role of cable to show feature films, and plans have been formulated for a joint French-German satellite venture, thereby raising questions about prospective demand for cable on the part of individual households who are all supposed to be linked into the nationwide network. Finally, notwithstanding the technological capabilities of optical fiber, chosen by the

French government over coaxial cable, it is not evident that a cable network will have economic advantages over other alternatives, such as a combination of satellite and terrestrial microwaves that has been used in the United States.

## II. Government Support to Industry: *R&D and Industrial Finance*

Soon after the socialist government came to power, spending on research and development was targeted to grow 20 percent a year between 1981 and 1985, and to increase from 2.0 to 2.5 percent of the gross domestic product (*GDP*) during this period. In the event, the share of *R&D* spending in *GDP* is estimated to have reached 2.2 percent of *GDP* as budgetary appropriations for research have been cut back in the period of austerity. The cutbacks affected primarily the financing of research programs that promote industrial research and provide equipment for laboratories. In turn, funds for the construction of the giant museum of science and technology at La Villette were not reduced.

France has traditionally practiced a policy of selective credit. This policy has been extended further after May 1981 through the establishment of CODEVI (*compte de développement industriel*) that provides advantages to savers over alternative instruments; the creation of FIM (*Fonds Industriel de Modernisation*) that receives an important part of CODEVI's resources for distribution to industrial firms and also borrows in domestic and in international markets; and the introduction of special institutional arrangements in favor of small- and medium-size enterprises.

The new institutions perform some of the functions customarily assigned to banks, in the present case the nationalized banks. Their operation has involved increased public interventions as loans by FIM are subject to approval by the ministry for foreign trade and industry, and other credit arrangements depend on decisions by local authorities. Furthermore, with the state guaranteeing loans by FIM in full, unprofitable operations may also receive financing. More generally, increasing the scope of selective credits tends

to reduce the efficiency of the allocation of financial resources while adding to the complexity of the credit system.

A further question is if the new institutions have actually increased the amount of funds available to industry. While data availabilities do not permit one to gauge the amount of credits accorded to the individual sectors, all enterprises taken together have seen their share in total credit decline as public deficits have increased. Thus, as Renaud de la Geniere, the Governor of the Banque de France until November 1984, noted in a presentation made at the Académie des Sciences Morales et Politiques on January 26, 1984, the net credit requirements of the public authorities that were practically nil in 1979 but reached 115 billion francs in 1983, compared with 160 billion francs for the enterprises whose net borrowings were 105 billion francs four years earlier. During the same period, the capacity of financing by households and financial institutions increased only from 105 to 200 billion francs, thereby necessitating borrowing abroad.

If the lack of credit is not an obstacle to investment (according to the latest survey by INSEE only 9 percent of industrial enterprises regarded credit to be a constraint), this seems less to do with the existence of new financial institutions as with factors such as insufficient profit margins (48 percent of responses) and the desire to avoid further increases in indebtedness (37 percent), which rose to a considerable extent in 1981 and 1982 when profit margins were particularly low. And while profit margins improved in 1983, the decline in industrial investment was not reversed.

With further improvements in profit margins, increases in industrial investment in 1984 and in 1985 may permit regaining the level reached in 1980, although public investment in steel accounts for a substantial part of the total. At the same time, the fall in industrial employment would continue and may even accelerate. Thus, annual decreases of 3 percent are estimated for 1984 and 1985, compared with declines of 2 percent in 1982 and in 1983.

The expected fall in industrial employment reflects the labor-saving character of

investments in response to higher labor costs, increased social charges, and, in particular, the difficulties encountered in attempting to reduce the size of the labor force if and when economic conditions warrant. Thus, as the freedom of enterprises to dismiss workers has been increasingly constrained, there has been a growing reluctance to hire labor, notwithstanding the measures taken in favor of the establishment of new enterprises in general, and small- and medium-size firms in particular. By contrast, greater flexibility in labor remuneration as well as in firings has contributed to the rapid expansion of employment in the United States.

### III. Concluding Remarks

In line with its expansionary stance, after May 1981 the socialist government postponed the necessary adjustment in French industries while wages and social charges increased to a considerable extent. And, once the need for adjustment came to be understood, the government wished to avoid firing superfluous workers in the nationalized firms. The financial situation of these firms was further aggravated by the interest costs of borrowing to finance their deficit over and above their increased budgetary allocation.

Thus, while budgetary allocations to the newly nationalized firms doubled (from 21.8 billion francs in 1980 to 43.4 billion francs in 1983), their deficit was nearly five times higher in 1983 (11.4 billion francs) than in 1980 (2.4 billion francs). During the same period, a surplus of 1.5 billion francs turned into a deficit of a similar magnitude, rising further to an estimated 9.0 billion francs in 1984 at Renault that has suffered the consequences of higher wages and social charges, price control, and social troubles fomented largely by the C.G.T.

At the same time, providing for the increased financial needs of the declining sectors has limited the availability of funds for the expansion of high-technology activities. Also, higher wages and social charges and price control discouraged investment activity in the years 1981–83. Correspondingly, new activities did not expand sufficiently to take the place of the old ones and industrial em-

ployment declined. A continuation of this trend is expected, notwithstanding a rise in industrial investments that tend to be labor saving.

In these circumstances, the question arises if one can speak about an industrial policy in France. As an official of the French government suggested in private conversation, the measures applied in regard to the three declining sectors may be considered social policy rather than industrial policy. Nor can we speak of an industrial policy in regard to the automobile industry, where the state has repeatedly countenanced actions taken by the unions. And while the policies applied aim at encouraging the development of the electronics industry, questions arise about the efficacy of the measures applied.

But how about the role of the nationalized firms as instruments of industry policy? After detailed interventions under Chevènement's tenure as minister failed to bring the expected results, Fabius reduced the extent of interventions while the policies applied by Cresson are not yet clear.

At the same time, the procedures of appointment of the managers of the nationalized enterprises give rise to concern. Adding union representatives to the board of directors has led to the adoption of an accommodating stance towards union demands on the part of management, in particular at Renault. Also, the contracts of two chief executives who showed considerable independence, Raymond Lévy of Usinor and Daniel Deguen of the Crédit Commercial de France, were not renewed. At the same time, there are signs that political considerations increasingly enter into the choice of chief executives.

In any case, the appointment of a chief executive for two years does not provide a sufficiently long learning period and favors the adoption of a short-term horizon. In fact, indications are that the chief executives of nationalized companies in France increasingly tend to take a short-term view, with the

neglect of long-term considerations, so as to increase their chances for reappointment.

Under the present ownership structure, these shortcomings may be mitigated by lengthening the terms of appointment of the chief executives and changing the composition of the board of directors, so as to minimize the chances of political interference. In this connection, reference may be made to attempts in Hungary for increasing the independence of the managers of state enterprises.

As to industrial policy, increased emphasis would need to be given to horizontal measures that operate across-the-board as against vertical measures that pertain to individual industries. In particular, there would be need to eliminate price control and to reverse increases in social charges. It would further be desirable to ease the constraints imposed on small- and medium-size enterprises in the form of additional social charges and workers' participation. Finally, greater flexibility would need to be introduced in labor markets.

The described measures may be expected to contribute to increased industrial investment and employment without the need for vertical interventions that tend to lead to inefficiencies in resource allocation. Horizontal measures would also need to be given emphasis in the support of research and development in the form of tax benefits for the *R&D* activities of the firms that generate external economies.

At the same time, the health of French industry, and of the French economy in general, is contingent on the pursuit of appropriate macroeconomic policies. Due to the weakness of investment activity and continued borrowing abroad, the rising budget deficit has not created a credit crunch for enterprises so far. But, with the constraints on foreign borrowing, the sustained rise of investment by enterprises would require that the public authorities limit their encroachment on domestic financial markets.



# *ECONOMIC HISTORY: A NECESSARY THOUGH NOT SUFFICIENT CONDITION FOR AN ECONOMIST<sup>†</sup>*

## Maine and Texas

By KENNETH J. ARROW\*

Henry David Thoreau said about the newly invented telegraph: "They tell us that Maine can now communicate with Texas. But does Maine have anything to say to Texas?" I suppose the existence and location of this panel shows that California and Massachusetts, at least, have much to say to Texas. But, as Dante explained so long ago, we must interpret texts not merely literally, but also allegorically and even spiritually. Our assigned subject asks that Maine and Texas be interpreted as Economic Analysis and Economic History, though not necessarily respectively.

I use the term "analysis," rather than "theory," because I want to contrast the aims of history and those of social science. Permit me to use the changes in my own thinking as illustration. I always found history very interesting and read heavily. As a graduate student, I qualified in economic history, which was compulsory, and never felt it to be an intrusion, as later generations of students did. As a faculty member, I had to consider the role of economic history in the curriculum, as well as the problem of filling vacancies there. I took it for granted that history was a necessity but recognized the difficulty of finding scholars who were both economists and historians. This was in the early 1950's, before a new generation solved that supply problem. I took it for granted that the role of history in the educa-

tion of economists was as empirical evidence. It was a way of testing theories, on a par with contemporary empirical evidence. This view was reinforced by what was then the most significant use of history in economics, Simon Kuznets's development of long time-series on national income and its components, with its strong implications for economic development and for the consumption function:

A lecture by the historian, Leonard Krieger, changed my view of the nature of history. There was then much discussion of the role of social science in history in general. Krieger was regarded as among the historians most sympathetic to social science. But in his lecture he made clear that history could not be regarded as simply a branch of social science. Its aims were different. It sought to study the individual case, while social science aimed at general principles. Social science, whether economics or another, might indeed be useful and even vital in interpreting a past event. Certainly, psychological theories of different kinds have been used, possibly not always with the best results, in interpreting the behavior of political leaders. But the use was to illuminate the particular event. The aim of historical study as such was not simply to serve as a source of data from which to infer and to test social science generalizations.

Of course, this does not preclude the use of the data thrown up by historical investigation for the purposes of social-scientific analysis. The two modes of inquiry are complementary, not substitutes. But they are not identical.

Let me draw a very close analogy from the natural world, that of geology. The underlying laws of geology are nothing but the standard laws of physics and chemistry.

<sup>†</sup>*Discussants:* Albert Fishow, University of California-Berkeley; Donald N. McCloskey, University of Iowa; Gavin Wright, Stanford University.

\*Departments of Economics and Operations Research, Stanford University, Stanford, CA 94305.

There is, from the viewpoint of scientific generalizations, nothing peculiar to geology. That water wears down rocks, that heat sources in the interior of the Earth can produce great changes in the Earth's surface, that the energy of the Earth derives in major part but not entirely from solar radiation, that under temperature and pressure the materials of the Earth form new chemical combinations, these relations and others determine the entire course of the Earth's development. Further, examination of the history of the Earth could in principle yield evidence about the measurements of specific chemical and physical reactions. In practice, though, observation in the laboratory is so much more efficient that the evidence of geological observation has probably been of little use to the underlying sciences except by way of suggesting research problems.

But geology is in fact a flourishing subject, and much of its interest is in the specific historical event. What is the history of the Appalachians and the Himalayas? What have been the movements of the Indian subcontinent? Why does Hawaii have the shape it does? It is in good measure a study of the specific. It is indeed history and a fascinating one.

The example of geology illustrates a recurrent topic in economics. Is economics a subject like physics, true for all time, or are its laws historically conditioned? The importance of history was on the rise throughout the nineteenth century, just when the abstract economic theory of David Ricardo was developed. Ricardo's doctrines were much attacked by contemporaries for lack of historical understanding. His disciple, John Stuart Mill, made clear that the laws of distribution were indeed historically conditioned; the classical laws of value held only in an economy in which exchange was governed by markets. So, too, the theory of plate tectonics is historically conditioned. It is a valid statement about the Earth today and for a long period in the past. It could not have held when the Earth was sufficiently hot, and it may or may not be true of other planets.

Physics and chemistry have clearly been very useful to geology, interpreted as history.

What does standard economic theory have to contribute to economic history? It could fail on several grounds. It might be so overwhelmingly powerful a theory that history becomes uninteresting, merely a playing out of a well-defined script. It could be so wrong that it is an obstacle to understanding. W. J. Cunningham attacked Ricardo and Marshall about 1890 for interpreting rent in the Tudor period in Ricardian terms, not recognizing the differences in historical conditions.

The first obstacle, the power of the theory, is clearly not valid, though some economic theorists have spoken as if it were. In form, neoclassical theory is a statement of the implications of tastes, technology, and expectations for prices and quantities. (Other variations of economic theory have a similar form, though different content.) There is plenty of room for historical specificity in the conditions even if economic theory were more reliable than it is in drawing conclusions from them. There is nothing in economic theory which specifies that tastes remain unchanged, and a great deal of empirical knowledge about changes in technology. Indeed, it may be complained rather that economic theory does not sufficiently constrain historical determination, particularly when the data are not sufficient.

In fact, I think it would be widely accepted that using the ideas and approaches of economic theory has been useful in economic history. Perhaps most important, economic theory has raised new questions for history. It asks how economic institutions work in redirecting the flow of resources, not merely their intended workings. Our views of the relations between railroads and economic growth in the nineteenth century, of the diffusion of specific technological innovations such as the reaper, or of the economic consequences and functioning of slavery have all been seriously altered in ways that required new ways of thinking and suggested search for new kinds of evidence. Measuring the economic conditions of the masses of the population may have been driven by political aims as much as by modern welfare economics, but the appropriate measures and data have certainly been much clarified by the latter. I have already alluded to long time-

series on national income, a concept drawn from economic analysis, as a major contribution to our understanding of the past, both as to constancies and as to structural change.

But the example of national income analysis does remind us of a danger in the use of economic theory in economic history. There is a bias towards flattening out the particularities of the past. The more one uses categories drawn from the need to generalize, the less marked is the difference among the instantiations. This is not a logical consequence of the use of theoretical constructs. As already emphasized, each historical episode can in principle be interpreted as the application of general principles to unique contexts; but the bias drawn from theory is likely to be to emphasize generality at the expense of particularity. One is reminded of earlier modes of historical interpretation, in which every catastrophe was the workings of the hand of God. Consider also the many theories for the interpretation of myths, where they are all considered as illustrations of some general principle, whether it be Frazer's dying king, Muller's rising and setting sun, or Freud's Oedipus complex, or even the structuralist seeking of a universal form to myth. What is lost is the sense that myths are different; it is not true that when you've heard one, you've heard them all.

What about the uses of history in the development of economic analysis? There are many, but let me pick two, both alluded to earlier. One is simply the use of economic history as a source of empirical evidence for testing theories and estimating relations, what I referred to earlier as my naïve view on the role of history. It is far from exhausting the content of history, but it is certainly one of its uses. When an examination of long time-series shows that interest rates do not fully adjust for inflation, as Lawrence Summers has recently shown, the routine acceptance of the Fisher effect in analysis of contemporary conditions must surely be questioned. The regular patterns of consumption described by Engel's laws can be confirmed by historical shifts in industrial structure as well as by budget studies, incidentally making it hard to maintain simple models of homogeneous economic growth. Studies of past hy-

perinflations and their endings test theories of inflation. The historical analysis of individual business cycles was a live field after the pioneering work of Walt Rostow from 1946 to about 1960; it was neglected because of one-sided theories, first Keynesian and then monetarist, but I hope Peter Temin's work signals a revival. Such work is both history itself and a testing ground for the many relations which define cyclical fluctuations.

A second use of history in the development of economic analysis is a definition of its historical conditioning, to pick up again an earlier point. Before economic analysis had much of an effect on economic history, historians debated whether and which earlier periods could be described as capitalist or almost so. The great classical historian, M. I. Rostovtzeff, found the early Roman Empire to be governed by modern economic institutions, mobility, profit seeking, and so forth. He has been ridiculed for this by the current leading authority, Moses Finley, who finds little evidence of rational economic behavior in the ancient world. Again, Henri Pirenne found merchants and traders in a few centuries around the year 1000 to be thoroughgoing profit seekers, acutely sensitive to price differences; but the crystallization of the guild system, according to Pirenne, subsequently created a different economic world. It is not for the theorists to assess the very specialized evidence available. Still, there is some suggestion that the economic world of the past is not entirely different from that of our theories in the sharp rise in real wages after the radical shift in the land-labor ratio occasioned by the Black Death.

Closely intertwined with historical conditioning of theory is national or cultural conditioning. The study of the past is similar to that of the present elsewhere. There are, or at least seem to be, very large differences even among capitalist countries in such basic economic variables as savings rates and rates and levels of productivity growth. We find them even between such economically and culturally similar nations as the United States and Canada; there are significant differences in per capita incomes, productivity (even controlling for capital equipment, as in the

automobile industry), and labor union membership and activity. When we go farther afield, the differences increase. Studies have shown, for example, that real wage flexibility is distinctly lower in Western Europe than in the United States. Among the many differences between Japan and the United States, we might notice differences in industrial organization even in virtually identical technologies. Large corporations there perform innovative tasks which here seem best done by small firms. Perhaps connected are the well-known or at least widely alleged differences in organizational loyalty and the role of consensus in decision making within the large firm.

Cross-country comparisons of this kind are analogous to historical comparisons in exploring the range of validity of economic generalizations. But the relation is deeper.

Presumably, the international differences, insofar as they are not simply explainable by differences in natural resources, are themselves the result of history. It has frequently been suggested by political scientists that the fact that the United States was created by a revolution while English-speaking Canada was in part a reaction to the American Revolution has important and lasting effects. The cultural differences between nations, with all their implications for polity and economy, are precipitates of past events, sometimes from the far past. In an ideal theory, perhaps, the whole influence of the past would be summed up in observations on the present. But such a theory cannot be stated in any complex uncontrolled system, not even for the Earth, as we have seen. It will always be true that practical understanding of the present will require knowledge of the past.

## Is History Stranger than Theory? The Origin of Telephone Separations

By PETER TEMIN AND GEOFFREY PETERS\*

The integrated Bell Telephone System that we all grew up with vanished on January 1, 1984, to be replaced by a new, more open telephone network still in the process of definition. AT&T's divestiture of the telephone-operating companies resulted from the settlement of an antitrust suit against AT&T begun in 1974 and carried forward under three administrations. The trial was conducted in 1981, largely under the direction of William Baxter, the first Assistant Attorney General for Antitrust in the Reagan Administration, and the settlement was negotiated by him and the management of AT&T.

Baxter articulated his theory of the case to Senator Thurmond's Judiciary Committee midway through the AT&T trial. He said, among other things, "If one argues for divestiture, one argues that the cross-subsidy problem is terribly important, that the vertical integration economies probably are not very great, and that regulatory supervision is unwanted and more deregulation is possible" (1981, p. 27). Divestiture, in other words, would solve "the cross-subsidy problem" without doing much violence to the telephone network.

What is the cross-subsidy problem? As expressed by the government's economics experts in its antitrust suit shortly before Baxter's testimony, it was that AT&T's integrated structure gave it the opportunity and the incentive to subsidize its competitive long-distance services with revenues from its regulated local monopolies. (Bruce Owen, 1981; Nina Cornell, 1981) This, presumably, was one of the antitrust violations on which the government's suit was based. It is en-

tirely in accord with the lines of the negotiated divestiture. And as expressed by the government economics experts, it was a theoretical proposition.

Yet Baxter said to the Judiciary Committee, "Historically, AT&T has subsidized local telephone service with longlines revenues" (p. 27). And he continued that this pattern "probably" continued at the time of his appearance before the committee. Historical observation, in other words, opposed the theoretical proposition. Confusion on this scale—in which the Assistant Attorney General took one position while his expert witnesses took the opposite—cries for explanation. The explanation demonstrates the need for historical as well as theoretical analysis.

A little terminology will facilitate the historical narrative. A telephone "station" is what we colloquially refer to as a telephone. A toll "board" is the local exchange switch which routes local calls within the exchange and toll calls out of the exchange. There are two distinct models of telephone communication which have given rise to two different modes of accounting. In station-to-station accounting, a long-distance call is thought of as going from one telephone (station) to another. In the board-to-board model, the same call is broken into parts. The parts between the individual stations and their local exchanges (boards) are considered local; the long-distance call goes only between the local exchanges, that is, from board to board.

AT&T established its first toll rate schedule in 1889, the rates becoming applicable as service was opened. AT&T established its accounting on a board-to-board basis, reflecting both its initial conception of telephone communication as it developed from local service to a national network and its corporate organization (FCC, pp. 370–75).

The Postmaster General operated the telephone system during World War I. Not sub-

\*Massachusetts Institute of Technology, Cambridge, MA 02139, and New England School of Law, Boston, MA 02116. This research was supported by a grant from AT&T. All interpretations and conclusions are ours alone.

ject to the jurisdictional limitations that would later separate state and federal regulatory authorities, he set uniform interstate and intrastate toll rates which remained in effect through 1926, when AT&T instituted an interstate rate reduction. The FCC negotiated four additional reductions in interstate telephone rates before World War II, but did not attempt to separate interstate and intrastate assets during these rate negotiations.

The issue of accounting models arose in a case of intrastate rate determination. Illinois Bell had contested a rate reduction ordered by the Illinois Commerce Commission in 1923, and the case reached the U.S. Supreme Court as *Smith v. Illinois Bell*. The Supreme Court ruled that the issue of whether the mandated rates were "confiscatory" under the Fourteenth Amendment could not be decided without "specific findings" on the allocation of Illinois Bell's assets between interstate and intrastate service.

Referring to the "indisputable fact" that "exchange property" is used both for intrastate and interstate service, the Court said that, "It is obvious that, unless an apportionment is made, the intrastate service to which the exchange property is allocated will bear an undue burden—to what extent is a matter of controversy" (*Smith v. Illinois Bell*, p. 151). In other words, the "specific findings" needed to determine whether rates were confiscatory were to be based on station-to-station accounting; intrastate telephone rates had only to be set high enough to earn a satisfactory return on the capital under state jurisdiction.

The issue was considered by various states in the 1930's, but the FCC only began to examine the division of assets between interstate and intrastate activities during World War II when wartime traffic raised Long Lines' profits above the level allowed by the FCC. The FCC needed both to reduce Long Line's profits and to respond to the federal government's needs during World War II, which included strenuous demands on the interstate telephone network. And the Commission was acutely aware of the concerns of state regulators about the difference between interstate and intrastate rates for telephone calls over the same distance. As interstate

rates declined, comparable toll calls within states came to cost more than calls crossing state lines (NARUC and FCC, 1951).

The traditional way of reducing Long Line's "excessive" profits, reducing interstate rates, therefore was doubly problematical: it would increase the disparity between interstate and intrastate toll rates and also the quantity of interstate toll service demanded by the nonmilitary public. The FCC therefore sought to reduce Long Lines' profits by moving some expenses of local telephone service into interstate jurisdiction, that is, by using station-to-station accounting. Agreement on the use of separations procedures that divided expenses along station-to-station lines was reached in a series of meetings between representatives of AT&T and several FCC Commissioners in January, 1943. (See our forthcoming article which contains documentation for this and following points.)

The wartime agreement on separations procedures was embodied in the 1947 *Manual of Separations*. The FCC refrained from endorsing the manual, but said it would not object to its use. Telephone rates were set on the basis that non-traffic-sensitive exchange capital was allocated to interstate toll service according to the relative use of telephones as measured by the "subscriber line use" (*SLU*), where  $SLU = (\text{minutes of interstate use}) / (\text{total minutes of use})$ .

AT&T acceptance of the new accounting and separations procedures undoubtedly was based on its appreciation of the changed regulatory environment arising from the establishment of the FCC. Before 1934, interstate telephone service was essentially unregulated, and it made sense to keep expenses in state jurisdiction and revenues out of it. As the FCC gained influence and AT&T's allowable interstate rate of return declined, AT&T lost the incentive to keep its rate base in state jurisdiction.

AT&T therefore acceded to the federal government's wartime needs and the state regulators' needs for toll-rate uniformity. After all, the structure of telephone rates was not a major issue in a unitary system. Rates henceforth would be based on station-to-station accounting, but AT&T preserved its internal board-to-board accounting, using sep-

arations procedures to convert the figures from one model to the other.

In response to a 1950 FCC inquiry into interstate rates, AT&T proposed an alteration of the 1947 *Manual* that would have shifted more of the local telephone plant into the interstate rate base, and avoided or moderated a fall in interstate rates. The national organization of state regulatory commissioners (NARUC) strongly supported the plan since it would have provided the opportunity to reduce intrastate rates. The FCC however rejected it as being inconsistent with *Smith v. Illinois Bell* on the grounds "that its adoption would have the effect of introducing an arbitrary method whereby interstate services subject to Federal jurisdiction would, in effect, be subsidizing services beyond that jurisdiction." The FCC seems to have argued that *SLU* was the "correct" way to allocate costs between local and toll services.

The NARUC sought redress through Senator Earnest W. McFarland, the Republican majority leader and chairman of the Communications Subcommittee of the Senate Interstate Commerce Committee. It appealed to him at its 1950 convention in Phoenix, Arizona, the senator's home state. He responded with a letter to the FCC in which he expressed his dismay at the Commission's apparent willingness to "shift the load from the big user to the little user; from the large national corporations which are heavy users of long distance to the average housewife and business or professional man who do not indulge in a great deal of long distance." Noting the growing disparity between interstate and intrastate toll rates for comparable calls, the senator said, "I am not in a position to pass upon the question as to whether the remedy suggested by NARUC is the proper one but I am certain that something should be done—and at once."

The FCC either misjudged or ignored the intent of Senator McFarland's letter in its reply. It sent him a long summary of the history of separations replete with facts and figures in which it characterized the disparity in toll rates as "natural," and asserted that it was fulfilling its legal mandate to regulate interstate rates only. Senator McFarland re-

plied sharply: "I believe that the Commission's six-page reply takes a strictly technical attitude toward the whole problem rather than the broad, constructive viewpoint required by the Communications Act.... Frankly, the Commission's reply is disappointing to me and to my colleagues whose interest and concern occasioned my original letter to you."

The FCC promptly reopened negotiations with AT&T resulting in a revision of separations procedures that shifted enough revenue requirements to interstate operations to justify two interstate rate increases, not decreases, in the next two years. These were not only the first interstate rate increases granted since the FCC was created, they also took place at a time when the trend of technology was reducing the cost of long-distance service. The toll rate disparity to which Senator McFarland had directed the FCC's attention was sharply reduced. The FCC enforced cost allocations in which—by its own admission—long-distance revenues were used to subsidize local service.

When Long Lines' earnings rose in 1955, the FCC negotiated another revision of the 1947 *Manual* that followed the lines of the plan rejected by the FCC five years earlier. The new revision shifted even more revenue requirements into interstate jurisdiction. A series of further revisions continued the process of shifting exchange plant into the interstate arena, demonstrating the lasting imprint of the agreement between Congress, state and federal regulators, and AT&T reached in the early 1950's.

The issue of accounting models arose again in the 1960's and 1970's in reference to different problems and clothed in different language. As potential competitors appealed to the FCC for access to parts of the interstate telephone market, the FCC became concerned about cross subsidization *among* interstate services. The ensuing debate about cost allocations between various Bell services pitted two theories against each other. The FCC opted for a system based on fully distributed cost, where joint costs are allocated to different services according to relative use, as measured by an analogue of *SLU*. This

clearly is station-to-station accounting under a more modern name. AT&T favored the use of long-run incremental costs, an equally clear extension of the board-to-board model of long-distance service. The FCC rejected the use of long-run incremental costs as too subjective, but the preceding discussion of separations shows that the actual use of fully distributed costs is no less arbitrary and subject to manipulation. (William Baumol, 1971; Leland Johnson, 1982, pp. 16 and 34)

The divestiture also raised the question of accounting models, albeit implicitly, since the telephone network was split up along board-to-board lines. AT&T and the other interexchange carriers furnish board-to-board long-distance service, now called interLATA service, while the Regional Holding Companies and their operating companies supply local (intraLATA) service. While some Casandras saw the immediate end of separations procedures and consequent steep rises in local rates, Congress was up in arms over this now traditional issue, and the FCC has—so far—fallen once again line. Current controversies over access charges preserve in new bottles the old wine of disputes over accounting models.

Where, then, does the current confusion over cross subsidies come from? Only under the board-to-board model is there a clear subsidy from long-distance to local service. And, as has been shown here, the FCC has rejected the use of that model for forty years. Using a game-theoretic definition (Gerald Faulhaber, 1975), there is no cross subsidy from long-distance to local service under station-to-station accounting (unless the FCC's statement of thirty years ago that *SLU* represents the true station-to-station cost allocation is still valid today). (See our forthcoming article.) Even though divestiture has brought board-to-board accounting out of the closet once again, it has not clarified the difference between the two accounting models.

Why does the confusion still exist? Is it just confusion stemming from inadequate analysis? Or does it serve larger purposes?

Instead of grappling with these important questions, let us return to the relation of history to theory. The need for historical analysis to clarify the relation between theory and observation should be clear. The predictions of economic theory may fail to express historical reality for many reasons. And the historical record itself may be confused to the point where it is unclear which predictions have been fulfilled. Careful historical analysis can be avoided only at the economist's peril.

## REFERENCES

- Baumol, William J., "Testimony," FCC Docket No. 18128, Bell Exhibit 18, October 15, 1971.
- Baxter, William J., "Testimony," U.S. Senate, Committee on the Judiciary, 97th Congress, 1st Session, Hearings, "DOJ Oversight: U.S. v. AT&T," August 6, 1981.
- Cornell, Nina W., "Testimony," *U.S. v. AT&T Co.*, June 19, 1981.
- Faulhaber, Gerald R., "Cross-Subsidization: Pricing in Public Enterprises," *American Economic Review*, December 1975, 65, 966-77.
- Johnson, Leland L., *Competition and Cross Subsidization in the Telephone Industry*, Santa Monica: Rand Corporation, 1982.
- Owen, Bruce W., "Testimony," *U.S. v. AT&T Co.*, June 22, 1981.
- Temin, Peter and Geoffrey Peters, "Cross-Subsidization in the Telephone Network," *Willamette Law Review*, forthcoming.
- Federal Communications Commission (FCC), *Investigation of the Telephone Industry in the United States, 1939*, New York: Arno Press, 1974.
- National Association of Railroad and Utilities Commissioners (NARUC) and Federal Communications Commission (FCC), Telephone Toll Rates Subcommittee, *Message Toll Telephone Rates and Disparities*, Washington, 1951.
- Smith v. Illinois Bell Telephone Co.*, 282 U.S. 133 (1930).



# Economic History and Economics

By ROBERT M. SOLOW\*

I have in the back of my mind a picture of the sort of discipline economics ought to be—or at least the sort of discipline I wish it were. If economics were practiced in that way there would be nothing problematical about its reciprocal relationship with economic history. It would be pretty clear what it is that economic theory offers to economic history and what economic history offers to economic theory. I will try to describe what I mean below.

For better or worse, however, economics has gone down a different path, not the one I have in mind. One consequence, not the most important one, but the one that matters for this discussion, is that economic theory learns nothing from economic history, and economic history is as much corrupted as enriched by economic theory. I will come to that, too, later on.

You will notice that I am using strong language. I am prepared to admit right away that I may be dead wrong in my judgements. But there is no point in pussyfooting. Bluntness may lead to an interesting discussion. After all, no one would remember the old German Historical School if it were not for the famous *Methodenstreit*. Actually, no one remembers them anyway. (There must be a lesson in that.)

To get right down to it, I suspect that the attempt to construct economics as an axiomatically based hard science is doomed to fail. There are many partially overlapping reasons for believing this; but since that is not the topic under discussion today, I do not have to lay them out in an orderly way. I hope the following hodgepodge will convey what I mean.

A modern economy is a very complicated system. Since we cannot conduct controlled

experiments on its smaller parts, or even observe them in isolation, the classical hard-science devices for discriminating between competing hypotheses are closed to us. The main alternative device is the statistical analysis of historical time-series. But then another difficulty arises. The competing hypotheses are themselves complex and subtle. We know before we start that all of them, or at least many of them, are capable of fitting the data in a gross sort of way. Then, in order to make more refined distinctions, we need *long* time-series observed under *stationary* conditions.

Unfortunately, however, economics is a social science. It is subject to Damon Runyon's Law that nothing between human beings is more than three to one. To express the point more formally, much of what we observe cannot be treated as the realization of a stationary stochastic process without straining credulity. Moreover, all narrowly economic activity is embedded in a web of social institutions, customs, beliefs, and attitudes. Concrete outcomes are indubitably affected by these background factors, some of which change slowly and gradually, others erratically. As soon as time-series get long enough to offer hope of discriminating among complex hypotheses, the likelihood that they remain stationary dwindles away, and the noise level gets correspondingly high. Under these circumstances, a little cleverness and persistence can get you almost any result you want. I think that is why so few econometricians have ever been forced by the facts to abandon a firmly held belief. Indeed, some of Fortune's favorites have been known to write scores of empirical articles without once feeling obliged to report a result that contradicts their prior prejudices.

If I am anywhere near right about this, the interests of scientific economics would be better served by a more modest approach. There is enough for us to do without pretending to a degree of completeness and

\*Department of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139.

precision which we cannot deliver. To my way of thinking, the true functions of analytical economics are best described informally: to organize our necessarily incomplete perceptions about the economy, to see connections that the untutored eye would miss, to tell plausible—sometimes even convincing—causal stories with the help of a few central principles, and to make rough quantitative judgments about the consequences of economic policy and other exogenous events. In this scheme of things, the end product of economic analysis is likely to be a collection of models contingent on society's circumstances—on the historical context, you might say—and not a single monolithic model for all seasons.

I hope no one here will think that this low-key view of the nature of analytical economics is a license for loose thinking. Logical rigor is just as important in this scheme of things as it is in the more self-consciously scientific one. The same goes for econometric depth and sophistication, maybe even more so. I mentioned "rough" quantitative judgment a moment ago, but that was only to suggest that the best attainable, in macroeconomics anyway, is not likely to be precise, if we are honest with ourselves and others. It would be a useful principle that economists should actually believe the empirical assertions they make. That would require more discipline than most of us now exhibit, when many empirical papers seem more like virtuoso finger exercises than anything else. The case I am trying to make concerns the scope and ambitions of economic model building, not the intellectual and technical standards of model building.

I claimed earlier that the natural relation between economics and economic history would be clear and straightforward if only economics were practiced in the fashion I have just sketched. Now I had better say what I meant. If economists set themselves the task of modeling particular contingent social circumstances, with some sensitivity to context, it seems to me that they would provide exactly the interpretive help an economic historian needs. That kind of model is directly applicable in organizing a historical narrative, the more so to the extent that the

economist is conscious of the fact that different social contexts may call for different background assumptions and therefore for different models.

The other direction of influence, what economic history offers to that kind of economic theory, is more interesting. If the proper choice of a model depends on the institutional context—and it should—then economic history performs the nice function of widening the range of observation available to the theorist. Economic theory can only gain from being taught something about the range of possibilities in human societies. Few things should be more interesting to a civilized economic theorist than the opportunity to observe the interplay between social institutions and economic behavior over time and place.

I am going to illustrate by referring to the work of W. H. B. Court, not merely because his book *The Rise of the Midland Industries* was on A. P. Usher's reading list when I took his course in the late 1940's. I choose Court for no better reason than that I happened to run across an obituary article about him in the *Proceedings* of the British Academy for 1982. (Since Court died in 1971, Fate did seem to be playing a hand.)

Here, for instance, is an excerpt from Court's volume on *Coal* in the U.K.'s official history of World War II.

Observers who found the conduct of the mineworkers puzzling assumed that, in the normal way, a man who finds himself faced with the possibilities of higher earnings will be prepared to put out extra effort to obtain them. An assumption about the conduct of an individual is as a rule, however, also an assumption of some sort about the society in which he lives and of which he is a member. The individual's demand for income, his views upon the getting and spending of money, are usually formed by the part of society which he is most in touch with. For most men the social code, whatever it may be in their time and place, is something which they accept as given and take over with little demur or questioning. Before one can assume that a demand for additional income existed

on the coalfields and could easily translate itself into extra work, one has to ask whether the mining community had those standards or those habits. If it did not, and if it was unable to develop them in a short time, then even a rapid rise of wage rates might bring about no appreciable change in the working habits of the industry.

In his own methodological writing, Court made the point explicitly that men "living as they do in different societies...make their decisions according to different schemes of values and according to the habits and structures of the society they find themselves living in." Therefore an economic historian should be an "observer and re-creator of the codes, loyalties and organizations which men create and which are just as real to them as physical conditions." Add to that a command over two-stage least squares and you have the kind of economic historian from whom theorists have most to learn, if only they are willing to try. I have naturally lit on this passage about the labor market because that is the branch of theory I happen to be engaged in right now, but no doubt the thought would apply equally well to consumer spending or rivalry among firms. I must promise myself, before I lecture again on wage bargaining, to ask my students to read the chapters on "The Wage Bargain" and "The Concept of the Minimum" in Court's *British Economic History, 1870-1914: Commentary and Documents*. I wonder what they will make of it.

So much for the normative. If you read the same journals I do, you may have noticed that modern economics has an ambition and style rather different from those I have been advocating. My impression is that the best and brightest in the profession proceed as if economics is the physics of society. There is a single universally valid model of the world. It only needs to be applied. You could drop a modern economist from a time machine—a helicopter, maybe, like the one that drops the money—at any time, in any place, along with his or her personal computer; he or she could set up in business without even bothering to ask what time and which place. In a little while, the up-to-date economist

will have maximized a familiar-looking present-value integral, made a few familiar log-linear approximations, and run the obligatory familiar regression. The familiar coefficients will be poorly determined, but about one-twentieth of them will be significant at the 5 percent level, and the other nineteen do not have to be published. With a little judicious selection here and there, it will turn out that the data are just barely consistent with your thesis adviser's hypothesis that money is neutral (or nonneutral, take your choice) everywhere and always, modulo an information asymmetry, any old information asymmetry, don't worry, you'll think of one.

All right, so I exaggerate. You will recognize the kernel of truth. We are socialized to the belief that there is one true model and that it can be discovered or imposed if only you will make the proper assumptions and impute validity to econometric results that are transparently lacking in power.

Of course there are holdouts against this routine, bless their hearts.

As I inspect current work in economic history, I have the sinking feeling that a lot of it looks exactly like the kind of economic analysis I have just finished caricaturing: the same integrals, the same regressions, the same substitution of *t*-ratios for thought. Apart from anything else, it is no fun reading the stuff any more. Far from offering the economic theorist a widened range of perceptions, this sort of economic history gives back to the theorist the same routine gruel that the economic theorist gives to the historian. Why should I believe, when it is applied to thin eighteenth-century data, something that carries no conviction when it is done with more ample twentieth-century data?

The situation reminds me of a story I once heard told by an anthropologist who had spent some months recording the myths and legends of a group of Apache in New Mexico. One night, just before she was scheduled to end her field work and depart, the Indians said to her: We have been telling you our legends all these months—why don't you tell us one of yours? The anthropologist thought fast and then responded brilliantly by telling the Indians a version of the story of Beowulf.

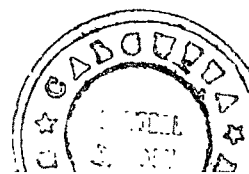
Years later she picked up a copy of an anthropological journal and found in the table of contents an article entitled "On the Occurrence of a Beowulf-like legend among the such-and-such Apache." If economic history turns into something that could be described as "The Occurrence of an Overlapping-Generations-like Legend among the Seventeenth-Century Neapolitans," then we are at the point where economics has nothing to learn from economic history but the bad habits it has taught to economic history.

Let me recapitulate. If the project of turning economics into a hard science could succeed, it would surely be worth doing. No doubt some of us should keep trying. If it did succeed, then there would be no difference between economics and economic history other than the source of data, no more than there is a difference between the study of astronomical events taking place now and those that took place in the Middle Ages. In this dispensation an economic historian is merely an economist with a high tolerance for dust or—what is rarer these days—a working knowledge of a foreign language.

There are, however, some reasons for pessimism about the project. Hard sciences dealing with complex systems—but possibly less complex than the U.S. economy—like the hydrogen atom or the optic nerve seem to succeed because they can isolate, they can experiment, and they can make repeated observations under controlled conditions. Other sciences, like astronomy, succeed because they can make long series of observations under natural but essentially stationary conditions, and because the forces being studied are not swamped by noise. Neither of these roads to success is open to economists.

In that case, we need a different approach. The function of the economist in this approach is still to make models and test them as best one can, but the models are more likely to be partial in scope and limited in applicability. "Testing" will have to be less mechanical and more opportunistic, encompassing a broader collection of techniques. One will have to recognize that the validity of an economic model may depend on the social context. What is here today may be gone tomorrow, or, if not tomorrow, then in ten or twenty years' time. In this dispensation there is a clear and productive division of labor between the economist and the economic historian. The economist is concerned with making and testing models of the economic world as it now is, or as we think it is. The economic historian can ask whether this or that story rings true when applied in earlier times or other places, and, if not, why not. So the economic historian can use the tools provided by the economist but will need, in addition, the ability to imagine how things might have been before they became as they now are. These are the sensitivities Court spoke of in the passage quoted above. I take it, naively perhaps, that they represent the comparative advantage of the historian.

In return, economic history can offer the economist a sense of the variety and flexibility of social arrangements and thus, in particular, a shot at understanding a little better the interaction of economic behavior and other social institutions. That strikes me as a meaningful division of labor. It was once suggested—by my kind of economist—that the division of labor is limited by the extent of the market. Perhaps what I have just been doing can be thought of as marketing.



# Clio and the Economics of QWERTY

By PAUL A. DAVID\*

Cicero demands of historians, first, that we tell true stories. I intend fully to perform my duty on this occasion, by giving you a homely piece of narrative economic history in which "one damn thing follows another." The main point of the story will become plain enough: it is sometimes not possible to uncover the logic (or illogic) of the world around us except by understanding how it got that way. A *path-dependent* sequence of economic changes is one of which important influences upon the eventual outcome can be exerted by temporally remote events, including happenings dominated by chance elements rather than systematic forces. Stochastic processes like that do not converge automatically to a fixed-point distribution of outcomes, and are called *non-ergodic*. In such circumstances "historical accidents" can neither be ignored, nor neatly quarantined for the purpose of economic analysis; the dynamic process itself takes on an *essentially historical* character. Standing alone, my story will be simply illustrative and does not establish how much of the world works this way. That is an open empirical issue and I would be presumptuous to claim to have settled it, or to instruct you in what to do about it. Let us just hope the tale proves mildly diverting for those waiting to be told if and why the study of economic history is a necessity in the making of economists.

\*Department of Economics, Encina Hall, Stanford University, Stanford, CA 94305. Support provided for this research, under a grant to the Technological Innovation Program of the Center for Economic Policy Research, Stanford University, is gratefully acknowledged. Douglas Puffert supplied able research assistance. Some, but not the whole, of my indebtedness to Brian Arthur's views on QWERTY and QWERTY-like subjects is recorded in the References. I bear full responsibility for errors of fact and interpretation, as well as for the peculiar opinions abbreviated herein. A fuller version with complete references, entitled "Understanding the Economics of QWERTY or Is History Necessary?," is available on request.

## I. The Story of QWERTY

Why does the topmost row of letters on your personal computer keyboard spell out QWERTYUIOP, rather than something else? We know that nothing in the engineering of computer terminals requires the awkward keyboard layout known today as "QWERTY," and we all are old enough to remember that QWERTY somehow has been handed down to us from the Age of Typewriters. Clearly nobody has been persuaded by the exhortations to discard QWERTY, which apostles of DSK (the Dvorak Simplified Keyboard) were issuing in trade publications such as *Computers and Automation* during the early 1970's. Why not? Devotees of the keyboard arrangement patented in 1932 by August Dvorak and W. L. Dealey have long held most of the world's records for speed typing. Moreover, during the 1940's U.S. Navy experiments had shown that the increased efficiency obtained with DSK would amortize the cost of retraining a group of typists within the first ten days of their subsequent full-time employment. Dvorak's death in 1975 released him from forty years of frustration with the world's stubborn rejection of his contribution; it came too soon for him to be solaced by the Apple IIC computer's built-in switch, which instantly converts its keyboard from QWERTY to virtual DSK, or to be further aggravated by doubts that the switch would not often be flicked.

If as Apple advertising copy now says, DSK "lets you type 20-40% faster," why did this superior design meet essentially the same rejection as the previous seven improvements on the QWERTY typewriter keyboard that were patented in the United States and Britain during the years 1909-24? Was it the result of customary, nonrational behavior by countless individuals socialized to carry on an antiquated technological tradition? Or, as Dvorak himself once suggested, had there

been a conspiracy among the members of the typewriter oligopoly to suppress an invention which they feared would so increase typewriter efficiency as ultimately to curtail the demand for their products? Or perhaps we should turn instead to the other popular "Devil Theory," and ask if political regulation and interference with the workings of a "free market" has been the cause of inefficient keyboard regimentation? Maybe it's all to be blamed on the public school system, like everything else that's awry?

You can already sense that these will not be the most promising lines along which to search for an economic understanding of QWERTY's present dominance. The agents engaged in production and purchase decisions in today's keyboard market are not the prisoners of custom, conspiracy, or state control. But while they are, as we now say, perfectly "free to choose," their behavior, nevertheless, is held fast in the grip of events long forgotten and shaped by circumstances in which neither they nor their interests figured. Like the great men of whom Tolstoy wrote in *War and Peace*, "(e) very action of theirs, that seems to them an act of their own free will, is in an historical sense not free at all, but in bondage to the whole course of previous history..." (Bk. IX, ch. 1).

This is a short story, however. So it begins only little more than a century ago, with the fifty-second man to invent the typewriter. Christopher Latham Sholes was a Milwaukee, Wisconsin printer by trade, and a mechanical tinkerer by inclination. Helped by his friends, Carlos Glidden and Samuel W. Soule, he had built a primitive writing machine for which a patent application was filed in October 1867. Many defects in the working of Sholes' "Type Writer" stood in the way of its immediate commercial introduction. Because the printing point was located underneath the paper carriage, it was quite invisible to the operator. "Non-visibility" remained an unfortunate feature of this and other up-stroke machines long after the flat paper carriage of the original design had been supplanted by arrangements closely resembling the modern continuous roller-platen. Consequently, the tendency of the typebars to clash and jam if struck in rapid

succession was a particularly serious defect. When a typebar stuck at or near the printing point, every succeeding stroke merely hammered the same impression onto the paper, resulting in a string of repeated letters that would be discovered only when the typist bothered to raise the carriage to inspect what had been printed.

Urged onward by the bullying optimism of James Densmore, the promoter-venture capitalist whom he had taken into the partnership in 1867, Sholes struggled for the next six years to perfect "the machine." From the inventor's trial-and-error rearrangements of the original model's alphabetical key ordering, in an effort to reduce the frequency of typebar clashes, there emerged a four-row, upper case keyboard approaching the modern QWERTY standard. In March 1873, Densmore succeeded in placing the manufacturing rights for the substantially transformed Sholes-Glidden "Type Writer" with E. Remington and Sons, the famous arms makers. Within the next few months QWERTY's evolution was virtually completed by Remington's mechanics. Their many modifications included some fine-tuning of the keyboard design in the course of which the "R" wound up in the place previously allotted to the period mark "." Thus were assembled into one row all the letters which a salesman would need to impress customers, by rapidly pecking out the brand name: TYPE WRITER

Despite this sales gimmick, the early commercial fortunes of the machine, with which chance had linked QWERTY's destiny remained terrifyingly precarious. The economic downturn of the 1870's was not the best of times in which to launch a novel piece of office equipment costing \$125, and by 1878, when Remington brought out its Improved Model Two (equipped with carriage shift key), the whole enterprise was teetering on the edge of bankruptcy. Consequently, even though sales began to pick up pace with the lifting of the depression and annual typewriter production climbed to 1200 units in 1881, the market position which QWERTY had acquired during the course of its early career was far from deeply entrenched; the entire stock of QWERTY-

embodying machines in the United States could not have much exceeded 5000 when the decade of the 1880's opened.

Nor was its future much protected by any compelling technological necessities. For, there were ways to make a typewriter without the up-stroke typebar mechanism that had called forth the QWERTY adaptation, and rival designs were appearing on the American scene. Not only were there typebar machines with "down-stroke" and "front-stroke" actions that afforded a visible printing point; the problem of typebar clashes could be circumvented by dispensing with typebars entirely, as young Thomas Edison had done in his 1872 patent for an electric print-wheel device which later became the basis for teletype machines. Lucien Stephen Crandall, the inventor of the second typewriter to reach the American market (in 1879) arranged the type on a cylindrical sleeve: the sleeve was made to revolve to the required letter and come down onto the printing-point, locking in place for correct alignment. (So much for the "revolutionary" character of the IBM 72/82's "golf ball" design.) Freed from the legacy of typebars, commercially successful typewriters such as the Hammond and the Blickensderfer first sported a keyboard arrangement which was more sensible than QWERTY. Then so-called "Ideal" keyboard placed the sequence DHIATENSOR in the home row, these being ten letters with which one may compose over 70 percent of the words in the English language.

The typewriter boom beginning in the 1880's thus witnessed a rapid proliferation of competitive designs, manufacturing companies, and keyboard arrangements rivalling the Sholes-Remington QWERTY. Yet, by the middle of the next decade, just when it had become evident that any micro-technological rationale for QWERTY's dominance was being removed by the progress of typewriter engineering, the U.S. industry was rapidly moving towards the standard of an upright front-stroke machine with a four-row QWERTY keyboard that was referred to as "the Universal." During the period 1895-1905, the main producers of non-typebar machines fell into line by offering "the Universal" as an option in place of the Ideal keyboard.

## II. Basic QWERTY-Nomics

To understand what had happened in the fateful interval of the 1890's, the economist must attend to the fact that typewriters were beginning to take their place as an element of a larger, rather complex system of production that was technically interrelated. In addition to the manufacturers and buyers of typewriting machines, this system involved typewriter operators and the variety of organizations (both private and public) that undertook to train people in such skills. Still more critical to the outcome was the fact that, in contrast to the hardware subsystems of which QWERTY or other keyboards were a part, the larger system of production was nobody's design. Rather like the proverbial Topsy, and much else in the history of economies besides, it "jes' grewed."

The advent of "touch" typing, a distinct advance over the four-finger hunt-and-peck method, came late in the 1880's and was critical, because this innovation was from its inception adapted to the Remington's QWERTY keyboard. Touch typing gave rise to three features of the evolving production system which were crucially important in causing QWERTY to become "locked in" as the dominant keyboard arrangement. These features were *technical interrelatedness*, *economies of scale*, and *quasi-irreversibility* of investment. They constitute the basic ingredients of what might be called QWERTY-nomics.

Technical interrelatedness, or the need for system compatibility between keyboard "hardware" and the "software" represented by the touch typist's memory of a particular arrangement of the keys, meant that the expected present value of a typewriter as an instrument of production was dependent upon the availability of compatible software created by typists' decisions as to the kind of keyboard they should learn. Prior to the growth of the personal market for typewriters, the purchasers of the hardware typically were business firms and therefore distinct from the owners of typing skills. Few incentives existed at the time, or later, for any one business to invest in providing its employees with a form of general human capital which so readily could be taken

elsewhere. (Notice that it was the wartime U.S. Navy, not your typical employer, that undertook the experiment of retraining typists on the Dvorak keyboard.) Nevertheless the purchase by a potential employer of a QWERTY keyboard conveyed a positive pecuniary externality to compatibly trained touch typists. To the degree to which this increased the likelihood that subsequent typists would choose to learn QWERTY, in preference to another method for which the stock of compatible hardware would not be so large, the overall user costs of a typewriting system based upon QWERTY (or any specific keyboard) would tend to *decrease* as it gained in acceptance relative to other systems. Essentially symmetrical conditions obtained in the market for instruction in touch typing.

These decreasing cost conditions—or *system scale economies*—had a number of consequences, among which undoubtedly the most important was the tendency for the process of intersystem competition to lead towards de facto standardization through the predominance of a single keyboard design. For analytical purposes, the matter can be simplified in the following way: suppose that buyers of typewriters uniformly were without inherent preferences concerning keyboards, and cared only about how the stock of touch typists was distributed among alternative specific keyboard styles. Suppose typists, on the other hand, were heterogeneous in their preferences for learning QWERTY-based “touch,” as opposed to other methods, but attentive also to the way the stock of machines was distributed according to keyboard styles. Then imagine the members of this heterogeneous population deciding in random order what kind of typing training to acquire. It may be seen that, with unbounded decreasing costs of selection, each stochastic decision in favor of QWERTY would raise the probability (but not guarantee) that the next selector would favor QWERTY. From the viewpoint of the formal theory of stochastic processes, what we are looking at now is equivalent to a generalized “Polya urn scheme.” In a simple scheme of that kind, an urn containing balls of various colors is sampled with replacement, and every drawing of a ball of a specified color results

in a second ball of the same color being returned to the urn; the probabilities that balls of specified colors will be added are therefore increasing (linear) functions of the proportions in which the respective colors are represented within the urn. A recent theorem due to W. Brian Arthur et al. (1983; 1985) allows us to say that when a generalized form of such a process (characterized by unbounded increasing returns) is extended indefinitely, the proportional share of one of the colors will, with probability one, converge to unity.

There may be many eligible candidates for supremacy, and from an *ex ante* vantage point we cannot say with corresponding certainty which among the contending colors—or rival keyboard arrangements—will be the one to gain eventual dominance. That part of the story is likely to be governed by “historical accidents,” which is to say, by the particular sequencing of choices made close to the beginning of the process. It is there that essentially random, transient factors are most likely to exert great leverage, as has been shown neatly by Arthur’s (1983) model of the dynamics of technological competition under increasing returns. Intuition suggests that if choices were made in a forward-looking way, rather than myopically on the basis of comparisons among the currently prevailing costs of different systems, the final outcome could be influenced strongly by expectations. A particular system could triumph over rivals merely because the purchasers of the software (and/or the hardware) expected that it would do so. This intuition seems to be supported by recent formal analyses by Michael Katz and Carl Shapiro (1983), and Ward Hanson (1984), of markets where purchasers of rival products benefit from externalities conditional upon the size of the compatible system or “network” with which they thereby become joined. Although the initial lead acquired by QWERTY through its association with the Remington was quantitatively very slender, when magnified by expectations it may well have been quite sufficient to guarantee that the industry eventually would lock in to a de facto QWERTY standard.

The occurrence of this “lock in” as early as the mid-1890’s does appear to have owed



something also to the high costs of software "conversion" and the resulting *quasi-irreversibility of investments* in specific touch-typing skills. Thus, as far as keyboard conversion costs were concerned, an important asymmetry had appeared between the software and the hardware components of the evolving system: the costs of typewriter software conversion were going up, whereas the costs of typewriter hardware conversion were coming down. While the novel, non-typebar technologies developed during the 1880's were freeing the keyboard from technical bondage to QWERTY, typewriter makers were by the same token freed from fixed-cost bondage to any particular keyboard arrangement. Non-QWERTY typewriter manufacturers seeking to expand market share could cheaply switch to achieve compatibility with the already existing stock of QWERTY-programmed typists, who could not. This, then, was a situation in which the precise details of timing in the developmental sequence had made it privately profitable in the short run to adapt machines to the habits of men (or to women, as was increasingly the case) rather than the other way around. And things have been that way ever since.

### III. Message

In place of a moral, I want to leave you with a message of faith and qualified hope. The story of QWERTY is a rather intriguing one for economists. Despite the presence of the sort of externalities that standard static analysis tells us would interfere with the achievement of the socially optimal degree of system compatibility, competition in the absence of perfect futures markets drove the industry prematurely into standardization *on the wrong system*—where decentralized decision making subsequently has sufficed to hold it. Outcomes of this kind are not so exotic. For such things to happen seems only too possible in the presence of strong technical interrelatedness, scale economies, and irreversibilities due to learning and habituation. They come as no surprise to readers prepared by Thorstein Veblen's classic passages in *Germany and the Industrial Revolution*

(1915), on the problem of Britain's undersized railway wagons and "the penalties of taking the lead" (see pp. 126–27); they may be painfully familiar to students who have been obliged to assimilate the details of deservedly less-renowned scribblings (see my 1971, 1975 studies) about the obstacles which ridge-and-furrow placed in the path of British farm mechanization, and the influence of remote events in nineteenth-century U.S. factor price history upon the subsequently emerging bias towards Hicks' labor-saving improvements in the production technology of certain branches of manufacturing.

I believe there are many more QWERTY worlds lying out there in the past, on the very edges of the modern economic analyst's tidy universe; worlds we do not yet fully perceive or understand, but whose influence, like that of dark stars, extends nonetheless to shape the visible orbits of our contemporary economic affairs. Most of the time I feel sure that the absorbing delights and quiet terrors of exploring QWERTY worlds will suffice to draw adventurous economists into the systematic study of essentially historical dynamic processes, and so will seduce them into the ways of economic history, and a better grasp of their subject.

### REFERENCES

- Arthur, W. Brian, "On Competing Technologies and Historical Small Events: The Dynamics of Choice Under Increasing Returns," Technological Innovation Program Workshop Paper, Department of Economics, Stanford University, November 1983.
- Arthur, W. Brian, Ermoliev, Yuri M. and Kaniovski, Yuri M., "On Generalized Urn Schemes of the Polya Kind," *Kibernetika*, No. 1, 1983, 19, 49–56 (translated from the Russian in *Cybernetics*, 1983, 19, 61–71).
- \_\_\_\_\_, \_\_\_\_\_, and \_\_\_\_\_, "Strong Laws for a Class of Path-Dependent Urn Processes," in *Proceedings of the International Conference on Stochastic Optimization*, Kiev, Munich: Springer-Verlag, 1985.
- David, Paul A., "The Landscape and the Machine: Technical Interrelatedness, Land Tenure and the Mechanization of the Corn Harvest in Victorian Britain," in D. N.

- McCloskey, ed., *Essays on a Mature Economy: Britain after 1840*, London: Methuen, 1971, ch. 5.
- , *Technical Choice, Innovation and Economic Growth: Essays on American and British Experience in the Nineteenth Century*, New York: Cambridge University Press, 1975.
- Hanson, Ward A., "Bandwagons and Orphans: Dynamic Pricing of Competing Technological Systems Subject to Decreasing Costs," Technological Innovation Program Workshop Paper, Department of Economics, Stanford University, January, 1984.
- Katz, Michael L. and Shapiro, Carl, "Network Externalities, Competition, and Compatibility," Woodrow Wilson School Discussion Paper in Economics No. 54, Princeton University, September, 1983.
- Veblen, Thorstein, *Imperial Germany and the Industrial Revolution*, New York: MacMillan, 1915.

## CREDIT AND ECONOMIC INSTABILITY<sup>†</sup>

### Portfolio Choice and the Debt-to-Income Relationship

By BENJAMIN M. FRIEDMAN\*

The relationship between outstanding debt and economic activity has attracted growing attention in recent years. In the United States, the principal focus of this attention thus far has been the empirical finding that, over time horizons ranging from a calendar quarter to a year or two, the total outstanding indebtedness of all U.S. borrowers other than financial intermediaries bears as close a relationship to income and prices as does any of the more familiar monetary aggregates or the monetary base. (See, for example, the evidence in my earlier paper, 1983.) This finding has potentially important implications for the conduct of monetary policy, and since 1983 the Federal Reserve System's semiannual reports to Congress have specified a growth range for "domestic nonfinancial credit" along with the growth ranges for three monetary aggregates.

A perhaps even more interesting aspect of the aggregate debt-to-income relationship in the United States is that the simple ratio of the U.S. economy's domestic nonfinancial debt to its gross national product has shown essentially no time trend over a period measured not in years but in decades. This finding bears potentially important implications for fiscal policy, especially in an era of federal budget deficits that are large enough to increase rapidly the federal government's out-

standing indebtedness in relation to *GNP*.<sup>1</sup> It is puzzling, however, in that many of the factors that familiar theory suggests would determine an economy's proclivity to finance its activity by issuing debt (for example, aggregate risk levels, tax rates, and bankruptcy arrangements) have changed dramatically over the decades during which the U.S. domestic nonfinancial debt ratio has remained approximately flat.

Any potential explanation for this phenomenon must, of course, focus on the behavior of lenders (debt holders) or borrowers (debt issuers), or both. The object of the research summarized in this paper is to see whether the behavior of investors in the U.S. financial markets could plausibly account for the economy's relatively stable debt-to-income ratio. This question turns on whether investors treat debt and other assets as close or distant substitutes in their portfolios. To anticipate, analysis of financial assets' respective risk properties indicates that debt and equity are indeed sufficiently distant substitutes for lenders' behavior to be a plausible source of the constraint confining the debt-to-income ratio within relatively narrow limits. At the same time, nothing in this finding precludes the possibility that borrowers' behavior could also be an equally or even more important part of the overall explanation.

#### I. Asset Risk and Asset Substitutability

The key link to lenders' behavior exploited here is the well-known fact that the U.S.

<sup>†</sup>*Discussants:* Hyman P. Minsky, Washington University-St. Louis; James S. Earley, University of California-Riverside.

\*Department of Economics, Harvard University, Cambridge, MA 02138. I am grateful to Jeff Fuhrer for research assistance; to him, Andrew Abel and James Earley for helpful comments on a preliminary draft; and to the National Science Foundation and the Alfred P. Sloan Foundation for research support.

<sup>1</sup>The federal government's debt ratio declined (as is usual in peacetime) from a peak of 1.03 in 1946 to a low of .25 in 1974. At year end 1980, it was still .27; by year end 1984, it had risen to .37.

economy's total wealth-to-income ratio has been essentially trendless for many decades (see Raymond Goldsmith, forthcoming), as would be implied by the life cycle model of saving under standard conditions describing a mature (albeit growing) economy. Over substantial periods of time, therefore, a stable debt-to-income ratio is equivalent to a stable share of debt assets in the economy's aggregate portfolio. In terms of familiar portfolio theory, if investors' behavior is imposing this constraint, then the relevant substitution elasticities must be small (in absolute value) in comparison with the corresponding wealth and/or income elasticities. Whether in fact they are so is an empirical question.

According to the standard theory describing the portfolio behavior of risk-averse investors, the relevant asset substitutabilities that matter here depend on investors' perceptions of the risk associated with holding debt and other assets. Investors' willingness to hold different assets depends on their assessments of the respective risks to which holding these assets exposes them, and their treatment of some assets as substitutes for others in their portfolios likewise depends on the relationships they perceive among the associated risks to holding these assets as well as others. If two assets expose holders to essentially the same set of risks, investors typically treat the two as close substitutes and allocate their portfolios accordingly. Assets subject to quite disparate risks are typically more distant substitutes, or perhaps even complements.

The basic framework of analysis used here is the familiar discrete-time theory relating risk-averse portfolio choice to expected asset returns. The investor's single-period objective, given initial wealth  $W_t$ , is to choose a vector of asset holding proportions  $\alpha_t$ , satisfying  $\alpha_t' \mathbf{1} = 1$ , to maximize expected utility  $E[U(\tilde{W}_{t+1})]$ . Under the conditions that  $U(W)$  is any power or logarithmic function (so that the Pratt-Arrow coefficient of relative risk aversion is constant), that the investor perceives the vector of real net asset returns  $\mathbf{r}_t$  to be distributed normally (or lognormally) with expectation  $\mathbf{r}_t^e$  and variance-covariance

structure  $\Omega_t$ , and that no available asset is riskless in real terms, solution of this problem yields

$$(1) \quad \alpha_t^* = B_t(\mathbf{r}_t^e + \mathbf{1}) + \pi_t,$$

where

$$(2) \quad B_t = \left\{ \frac{-U'[E(\tilde{W}_{t+1})]}{W_t \cdot U''[E(\tilde{W}_{t+1})]} \right\} \cdot [\Omega_t^{-1} - (\mathbf{1}'\Omega_t^{-1}\mathbf{1})^{-1}\Omega_t^{-1}\mathbf{1}\mathbf{1}'\Omega_t^{-1}];$$

$$(3) \quad \pi_t = (\mathbf{1}'\Omega_t^{-1}\mathbf{1})^{-1}\Omega_t^{-1}\mathbf{1}.$$

If the time unit is sufficiently small to render  $W_t$  a good approximation to  $E(\tilde{W}_{t+1})$  for purposes of the underlying expansion, then the first (scalar) term within brackets in (2) is simply the reciprocal of the constant coefficient of relative risk aversion.

Matrix  $B_t$  in (1), expressing the response of each proportional asset demand to movements in the expected real returns on that and other assets, contains the set of relative asset substitutabilities that determine how stable the respective shares of the typical investor's portfolio will be. The solution for  $B_t$  in (2) makes clear the central role of investors' risk perceptions in governing this behavior. The asset substitutabilities in  $B_t$  depend only on the investor's risk aversion and risk perceptions, here parameterized by a variance-covariance matrix  $\Omega_t$  that in general may vary over time.

## II. Substitutability among Financial Assets

Table 1, panel A, shows the variances and covariances, calculated from quarterly data for 1960–80, of the realized after-tax real per annum returns on three broad classes of U.S. financial assets that differ fundamentally from one another according to the risks associated with holding them. Short-term debt ( $S$ ) includes all assets bearing real returns that are risky, over a single year or calendar quarter, only because of uncertainty about inflation. Long-term debt ( $L$ ) is risky because of uncertainty not only about inflation but also about changes in asset prices di-

TABLE 1—ASSET RETURN RISKS AND IMPLIED PORTFOLIO RESPONSES

	$r_S$	$r_L$	$r_E$
A. Variance-Covariance Matrix			
$r_S$	11.18		
$r_L$	29.91	209.35	
$r_E$	30.24	161.77	597.86
B. Portfolio Response Matrix			
$\alpha_S$	.641		
$\alpha_L$	-.578	.727	
$\alpha_E$	-.0635	-.150	.213

Notes: Asset returns scaled in percent per annum. Portfolio responses based on relative risk aversion equal to one.

rectly reflecting changes in market interest rates. Equity ( $E$ ) is risky because of uncertainty about inflation and about changes in stock prices. (For details of the construction of these three after-tax real returns, see my 1984 paper.)

Table 1, panel B, indicates the implications of this observed 1960–80 covariance structure for investors' portfolio behavior by showing the transformation of  $\Omega$  given in (2), up to but not including multiplication by the reciprocal of the coefficient of relative risk aversion. Apart from the risk-aversion coefficient, these values for B indicate the marginal responses of the proportional portfolio allocations  $\alpha$  to changes in expected asset returns  $r^e$ . Hence they also indicate by what amount the structure of expected returns would have to change in order to induce any given shift in the composition of the typical investor's portfolio.

For plausible values of the risk aversion coefficient, the B values shown in Table 1 indicate that short- and long-term debt are fairly close substitutes for one another, but not for equity. For a relative risk-aversion coefficient of four,<sup>2</sup> for example, the increase in the expected short-term debt return (relative to the two other returns) that would be

TABLE 2—IMPLICATIONS OF CONTINUALLY UPDATED RETURN FORECASTING

	$r_S$	$r_L$	$r_E$
A. Variance-Covariance Matrix			
$r_S$	1.25		
$r_L$	3.62	76.61	
$r_E$	6.45	48.09	317.27
B. Portfolio Response Matrix			
$\alpha_S$	1.57		
$\alpha_L$	-1.41	1.61	
$\alpha_E$	-.161	-.204	.365

Notes: See Table 1.

required to raise the short-plus-long debt share of the typical investor's portfolio by .01 is .63 percent. The corresponding required increase in the expected long-term debt return is .27 percent. Because the model is linear in expected returns, the analogous increase required to generate greater portfolio shifts are proportionally greater.

One potentially serious shortcoming of drawing such inferences on the basis of an unconditional sample variance-covariance structure is that it attributes too little information to investors by disregarding their knowledge, at each point in time, of the most recent realizations of asset returns and their principal determinants. During the 1960–80 period the after-tax real returns on all three classes of assets considered here exhibited substantial serial correlation. When returns are serially correlated, information about the most recent actual values is a useful ingredient in forming expectations about returns in the immediate future. Ignoring that information can lead to excessively large estimates of the uncertainty surrounding these expectations.

Table 2 presents a set of analogous results based on a procedure that takes much more careful account of what information investors did and did not have at any particular time. As of the beginning of each calendar quarter, investors presumably know the stated interest rates on short-term debt instruments, the current prices and the coupon rates on long-term debt instruments, the current prices and (approximately) the dividends on equities, and the relevant tax

<sup>2</sup>This value is about in the middle of the range of available empirical estimates. Irwin Friend and Marshall Blume (1975) suggested a value in excess of two, Sanford Grossman and Robert Shiller (1981) suggested four, and Friend and Joel Hasbrouck (1982) suggested six.

rates. The three uncertain elements that they must forecast over the coming quarter, in order to form expectations of the after-tax real returns on the three broad classes of assets considered here, are inflation, the capital gain or loss due to changing bond prices, and the capital gain or loss due to changing stock prices.

The procedure underlying the results reported in Table 2 represents investors as forming expectations of these three uncertain return elements, at each point in time, by estimating a linear regression model relating each element to past values of itself and the other two, using all data observed through the immediately preceding period.<sup>3</sup> In addition to providing forecast values of the three uncertain elements for the period ahead, the linear regression model at each point in time also directly indicates the variances and covariances associated with the forecasts derived in this way. After each period elapses, investors can then repeat the same procedure, incorporating the one new observation on inflation and on long-term debt and equity capital gains into the data used to reestimate the linear regression model to make forecasts for the next period.

Given the simple arithmetic connection between asset returns and these underlying uncertain elements, and given investors' presumed knowledge of the other elements comprising returns, these one-period-ahead forecasts of inflation and the respective capital gains on long-term debt and equity directly imply one-period-ahead forecasts of the after-tax real returns on all three classes of assets at each point in time. Similarly, the variances and covariances associated with the forecasts of inflation and the two capital gains directly imply the variances and covariances associated with the corresponding forecasts of the three asset returns.

Table 2, panel A, shows the means of these implied return variances and covariances for the 84 quarters of the sample. These values are smaller than the corresponding values shown in Table 1, indicating the importance of investors' having (and

using) information about recent actual returns.

Table 2, panel B, shows the transformation of this  $\Omega$  given in (2), again up to but not including multiplication by the risk-aversion reciprocal. The reduced uncertainty, in comparison with Table 1, makes investors more willing to re-allocate their portfolios in response to changes in expected returns. Even so, most of the asset substitutability is still between short- and long-term debt. With relative risk aversion again equal to four, the increases in the expected returns on short- and long-term debt required (individually) to raise the overall debt share of the typical investor's portfolio by .01 are .25 and .20 percent, respectively.

### III. Substitutability between Financial and Nonfinancial Assets

An important limitation of the analysis reported in Section II is its restriction to financial assets only. On a net basis, most of the total U.S. national wealth that has remained relatively stable in relation to U.S. economic activity consists of nonfinancial assets. Even for the household sector alone, year-end 1980 total wealth included \$2.8 trillion of residential real estate and \$1.0 trillion of consumer durables in addition to \$3.5 trillion of financial assets. If wealth holders are willing to substitute not just among financial assets but also between financial and nonfinancial assets, then the results presented in Section II presumably overstate the movements in the expected return structure required to change the share of debt in their portfolios, and hence also overstate the likely resulting stability of aggregate debt holdings in relation to either wealth or income.

Table 3 presents the results of applying the forecasting procedure underlying Table 2 to the after-tax real returns on the same three classes of financial assets together with two classes of nonfinancial assets, residential real estate ( $H$ ) and consumer durables ( $D$ ), based on annual data for 1964–81.<sup>4</sup> Apart

<sup>3</sup>See my 1984 paper for details of the estimated vector autoregression and the calculations based on it.

<sup>4</sup>The nominal after-tax return for housing combines the BEA implicit rent and depreciation series, the FHA series on maintenance costs, the MPS series on property

TABLE 3—CONTINUALLY UPDATED RETURN  
FORECASTING INCLUDING NONFINANCIAL ASSETS

	$r_S$	$r_L$	$r_E$	$r_H$	$r_D$
A. Variance-Covariance Matrix					
$r_S$	2.31				
$r_L$	6.62	43.76			
$r_E$	11.92	49.92	191.03		
$r_H$	.47	1.24	4.47	1.19	
$r_D$	.67	1.87	2.99	.16	.27
B. Portfolio Response Matrix					
$\alpha_S$	176				
$\alpha_L$	-14.5	6.25			
$\alpha_E$	-4.49	-.628	.942		
$\alpha_H$	34.0	6.79	-2.95	168	
$\alpha_D$	-191	2.12	7.13	-206	387

Notes: See Table 1.

from the use of an annual time unit, the treatment of the uncertain elements of the financial asset returns is just analogous to that described in Section II. In order to generate forecasts of the respective returns on housing and durables, however, the forecasting equation here also includes the change in the constant-quality housing price index and the change in the implicit price deflator for durables.

Not surprisingly, given the role of inflation in making asset returns uncertain, the resulting variance-covariance matrix shown in panel A, Table 3, indicates that both categories of nonfinancial assets are less risky in real terms than any of the three financial assets. More importantly for the purposes of the analysis here, the transformation of this variance-covariance structure shown in panel B indicates that the implied responsiveness of portfolio allocations to changes in expected returns is far greater than suggested

by the analysis in Section II of financial assets alone. With relative risk aversion equal to four, the increase in the expected short-term debt return (again, relative to all other returns) required to raise the total debt share of the typical investor's portfolio by .01 is only .025 percent. A comparison of the elements in the first column of the matrix makes clear that more than all of this portfolio re-allocation occurs at the expense of the share invested in durables. Because of the cross effects of the substitutability of short-term debt with both durables and long-term debt, however, the corresponding movement in the expected long-term debt return required to raise the total debt share by .01 is a decline of .48 percent.

#### IV. Conclusions

Whether or not investors' behavior can plausibly account for the U.S. economy's trendless debt-to-income ratio depends crucially on the proper treatment of wealth holding in nonfinancial forms. Among financial assets only, the substitutability of debt and equity securities is sufficiently limited that very large movements in expected return differentials—movements so large as presumably to elicit offsetting responses from borrowers—would be required to induce major changes in the debt share of investors' aggregate portfolio. Given the long-run stability of the economy's wealth in relation to income, this lack of asset substitutability along the relevant dimension also implies a stable debt-to-income ratio.

By contrast, a parallel analysis applied to financial and nonfinancial assets together suggests that only quite modest movements in the structure of expected returns would suffice to induce even very large changes in the debt share of total assets, and hence in the aggregate debt-to-income ratio. The main reason for this result is the close substitutability of short-term debt and consumer durables implied by the respective risks associated with these two assets' after-tax real returns.

Especially since the key substitutability on which this difference hinges is to durables, rather than housing, the most sensible inter-

taxes, and changes in the Census Bureau constant-quality price index, using Robert Barro and Chaipat Sahasakul's (1983) average marginal income tax rate series. The nominal (untaxed) return on durables combines the BEA service value estimate and changes in the relevant BEA deflator. In both cases the corresponding real return follows from subtracting the percentage change in the consumer price index. Use of an annual time unit in this part of the analysis reflects the unavailability of several of these series on a quarterly basis.

pretation of these results is probably to discount the findings including nonfinancial assets and conclude that the portfolio behavior of risk-averse investors can plausibly account for a stable debt share of assets, and hence also (given the stability of wealth in relation to income in the United States) the observed stable debt-to-income ratio. One reason for drawing this conclusion is simply that asset-type considerations of risk and return alone probably do not constitute an adequate description of the demand for consumer durables. In a more fully developed description of that demand, the willingness to substitute holdings of short-term debt instruments for ownership of consumer durables would no doubt be much more limited. A second reason is that, to a far greater extent than in the case of return indexes for aggregates of financial assets (or even housing), the variation of the return index for the aggregate of all consumer durables presumably understates the risk associated with any individual's holding. A more accurate representation of that risk would also probably indicate less correlation with other asset risks, hence less substitutability for other assets, and hence less responsiveness of asset demands to changes in relative returns.

With this qualification, therefore, the behavior of lenders in the U.S. financial market does exhibit characteristics that could account for the observed stability of the economy's aggregate debt-to-income ratio over long periods of time. This conclusion, however, in no way precludes the behavior of

borrowers being as important, or more so, in explaining this phenomenon. That possibility remains a subject for future research.

## REFERENCES

- Barro, Robert J. and Sahasakul, Chaipat, "Measuring the Average Marginal Tax Rate from the Individual Income Tax," *Journal of Business*, October 1983, 56, 419-52.
- Friedman, Benjamin M., "The Roles of Money and Credit in Macroeconomic Analysis," in James Tobin, ed., *Macroeconomics, Prices and Quantities: Essays in Memory of Arthur M. Okun*, Washington: The Brookings Institution, 1983.
- \_\_\_\_\_, "Crowding Out or Crowding In? Evidence on Debt-Equity Substitutability," mimeo., National Bureau of Economic Research, 1984.
- Friend, Irwin and Blume, Marshall E., "The Demand for Risky Assets," *American Economic Review*, December 1975, 65, 900-22.
- \_\_\_\_\_, and Hasbrouck, Joel, "Effect of Inflation on the Profitability and Valuation of U.S. Corporations," in M. Sarnat and A. Szego, eds., *Savings, Investment and Capital Markets in an Inflationary Economy*, Cambridge: Ballinger, 1982.
- Goldsmith, Raymond W., *Comparative National Balance Sheets*, Chicago: University of Chicago Press, forthcoming.
- Grossman, Sanford J. and Shiller, Robert J., "The Determinants of the Variability of Stock Prices," *American Economic Review Proceedings*, May 1981, 71, 222-27.



# Stability and Instability in the Debt-Income Relationship

By ROBERT POLLIN\*

Over the recent past, economists have increasingly pursued research on the question of financial instability and financial crisis, reflecting, of course, the growing seriousness of these issues in the real world, both domestically and internationally. One important aspect of this research has been the effort to establish a set of empirical relationships through which tendencies toward instability and crisis may be accurately observed, and thus better understood. This paper has a dual purpose: first, to consider empirical measures of financial activity in the U.S. economy, specifically the trend relationship between nonfinancial debt and *GNP*; then, based on the empirical discussion, to offer an approach toward understanding some of the sources of contemporary financial instability. In the latter aim, I pay particular attention to the issue of federal government deficits.

## I. Stock and Flow Measures of Nonfinancial Debt

In studies of the U.S. economy, one empirical relationship that has attracted considerable interest is that between the outstanding debt of the economy's nonfinancial sectors and *GNP*—the juxtaposition of the nonfinancial economy's accumulated stock of net liabilities and its ability to generate income and output. This relationship has been investigated most thoroughly by Benjamin Friedman (1982, 1984), who has identified two important aspects of it. First, Friedman has found a high degree of stability between the growth of total outstanding debt of the nonfinancial sectors and *GNP*. As can be seen in Figure 1, this total outstanding debt-*GNP* ratio ( $S_t$ ) has displayed

essentially no trend and only limited cyclical variation throughout the post-World War II period. In contrast with this stability of the total outstanding debt ratio, Friedman has also found that the outstanding debt ratios of the individual nonfinancial sectors (households, businesses, and government) has varied significantly and in divergent ways. This also can be seen in Figure 1.<sup>1</sup> In particular, we observe a substantial declining trend of outstanding federal government debt-*GNP* ( $S_g$ ), and a mirroring of this by a rising trend of the outstanding debt-*GNP* ratios for households ( $S_h$ ) and nonfinancial business corporations ( $S_c$ ).

Friedman correctly recognizes that, as yet, no adequate explanation of these empirical patterns has been formulated. Nevertheless, a useful place at which an analysis may begin is the pioneering effort of John Gurley and Edward Shaw (1957) to explain the long-term tendency toward stability of the total outstanding debt ratio  $S_t$  (they did not consider the movements of its component parts). Gurley and Shaw formulated a long-term growth model of a developing economy, beginning at the zero level for both *GNP* and outstanding debt. They made two key assumptions in the model: a stable growth rate of *GNP* (after controlling for short cycles), and a constant marginal propensity to issue net new debt relative to *GNP*. Based upon their assumptions, they showed that  $S_t$  in their hypothetical economy does become asymptotically stable as the economy matures. The formula they derived for producing this stable ratio in a mature economy is

$$(1) \quad S_t = f_t(1 + y)/y,$$

where  $y$  is the growth rate of *GNP* and  $f_t$  is the marginal propensity of the aggregate

\*Department of Economics, University of California, Riverside, CA 92521. I thank James Earley, Benjamin Friedman, Mason Gaffney, Albert Wojnilower, and Martin Wolfson for insightful comments; Ted Schmidt and Timothy Tracy for research assistance; and the U.C.-Riverside Committee on Research for financial support.

<sup>1</sup>In Figure 1, and the figures following, the sectoral ratios for state and local government and unincorporated business have been omitted.

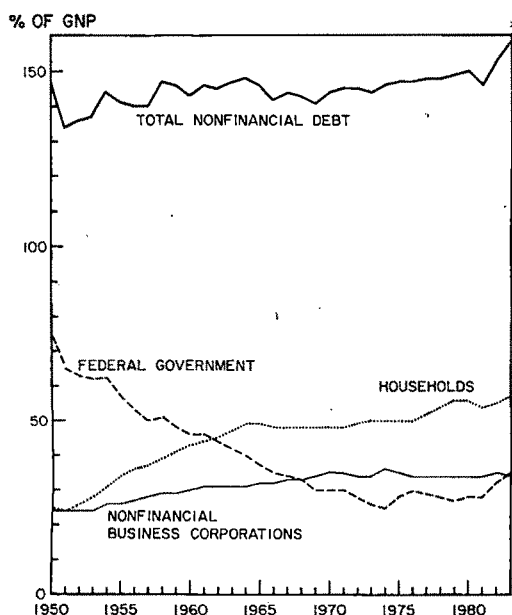


FIGURE 1. OUTSTANDING DEBT RATIOS FOR U.S. NONFINANCIAL BORROWERS

nonfinancial sector to issue net new debt—this latter variable I specifically define as net aggregate debt *flows*-GNP and refer to as the aggregate “debt-financing” ratio. According to Gurley and Shaw, therefore, the basis for the stability of  $S_t$  in a mature economy is the underlying stability which they assumed for  $f_t$  and  $y$ .

What is remarkable about the actual stability of  $S_t$  throughout the postwar period, and especially since the 1960's, is that it has resulted not through conformity with the

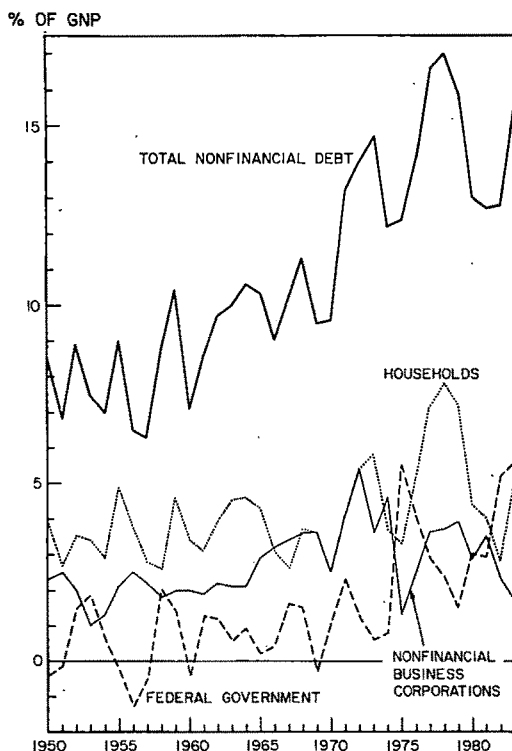


FIGURE 2. DEBT-FINANCING RATIOS FOR U.S. NONFINANCIAL BORROWERS

Gurley-Shaw conditions (a stable  $f_t$  and  $y$ ) but, in fact, *despite the sustained violation of these conditions*. First, GNP growth has not been stable over the postwar period on a trend basis, neither in real or nominal terms. Rather, the growth trend for nominal GNP has been upward throughout the period, while with real GNP growth, the trend was upward from the mid-1950's to mid-1960's and downward thereafter. These patterns can be seen in Table 1, where data are grouped by cycles. In addition, and even more significant for present concerns, the aggregate debt-financing ratio  $f_t$  has also not been stable over the full postwar period. Until the mid-1960's,  $f_t$  was trendless, but since then it has risen sharply on a trend basis. This pattern can be seen on the top line of Figure 2. The lower lines of Figure 2 show that each of the three major borrowing sectors (the federal government, households, and nonfinancial corporations) have also experienced rising

TABLE 1—POSTWAR GNP GROWTH RATES

Cycles	Nominal Growth	Real Growth
1949-51	8.6	5.8
1952-55	4.9	3.2
1956-59	5.1	2.4
1960-66	6.5	4.6
1967-73	8.4	3.5
1973-79	10.5	2.8
1980-83	8.2	0.9

Note: Cycles are measured according to the ratio of actual to potential GNP, the “GNP gap” ratio. The rates are shown in percent.

debt-financing ratios on a trend basis since the mid-1960's. And while we do also observe inverse covariation here between the federal government's debt-financing ratio ( $f_g$ ) and those for households ( $f_h$ ) and non-financial corporations ( $f_c$ ), in this case the relationship holds only on a cyclical basis, not over the period as a whole.

Hence, the actual stability of the total outstanding debt ratio  $S_t$  since the 1960's has emerged out of a set of underlying phenomena quite contrary to those posited by Gurley and Shaw: that is, a stable  $S_t$  has resulted through a declining trend for real  $y$  and rising trends for nominal  $y$  and, most importantly,  $f_t$  (all debt ratios being calculated with nominal  $GNP$  as the denominator). In other words, considering the trends for  $S_t$  and  $f_t$ , the aggregate nonfinancial sector has been accumulating debt since the mid-1960's at a stable rate relative to  $GNP$  while, by contrast, the aggregate rate of debt financing has been rising. At the individual sectoral level, moreover, we observe a similar divergence between outstanding debt and debt-financing ratios:  $S_g$  moves in inverse covariation with  $S_h$  and  $S_c$ , while, on a trend basis,  $f_g$ ,  $f_h$ , and  $f_c$  vary directly. To obtain a clear understanding of financial market behavior in the contemporary U.S. economy, and particularly its tendencies toward instability, it seems apparent that these contrary patterns for the outstanding debt and debt-financing ratios need to be examined carefully.

## II. Reconciling the Divergent Trends

To explain the divergent patterns of  $S_t$  and  $f_t$ , there is a simple formal explanation. Referring to the Gurley-Shaw equation  $S_t = f_t(1+y)/y$ , it is evident that since  $y$ , the nominal growth rate of  $GNP$ , has been rising sharply since the mid-1960's, the ratio  $(1+y)/y$  will fall correspondingly over this period. Therefore  $f_t$  must rise along with  $y$  in order for  $S_t$  to remain constant. But this formal solution does not explicitly address the crucial substantive phenomenon affecting the divergent patterns of  $S_t$  and  $f_t$ —the asymmetric impact of inflation on the two ratios. Consider first  $S_t$ . Here the numerator

of the ratio, the stock of debt, remains fixed in nominal terms regardless of variations in the price level. The denominator, nominal  $GNP$ , is a flow variable however, and as such it varies in nominal terms directly with the price level. In an inflationary environment therefore, the nominal value of the debt stock remains fixed while  $GNP$  rises, so that  $S_t$  is biased downward. With  $f_t$ , by contrast, current-period flow values are in both numerator and denominator, and thus the impact of inflation on the ratio is neutral. It is therefore because of the sustained inflation since the 1960's that  $S_t$  has remained stable while  $f_t$  has risen.<sup>2</sup>

To explain the inverse covariation of  $S_g$  with both  $S_h$  and  $S_c$ , and the absence of similar trend relationships with the sectoral debt-financing ratios, we must focus upon financial conditions emerging out of World War II. During the war, the federal government accumulated a huge debt while concurrently restrictions were placed on private sector expenditure and borrowing. As a result, the government owed 71 percent of the total U.S. outstanding debt in 1945. With this figure as the starting point for the postwar period, a pattern of inverse covariation between  $S_g$  and both  $S_h$  and  $S_c$  necessarily emerges after the war as long as  $f_g$  is less than  $f_h$  and  $f_c$  (a condition which holds for virtually every postwar year until the mid-1970's). This inverse covariation will then necessarily continue until  $S_g$  settles at a level that reflects the government's current proportion of net new debt issues—that is, until the effects of wartime finance on the outstanding debt ratios have dissipated.

To illustrate this point further, Figure 3 shows the results of the following simulation exercise: the postwar period begins in 1946 with the actual proportions of debt outstand-

<sup>2</sup>To recognize this effect of inflation on debt ratios does not however point toward any particular causal hypothesis explaining the relationship between the rates of inflation and debt issuance since the mid-1960's. This complex question still requires much further investigation. I can however acknowledge here an important related accounting effect of inflation. See my 1984 paper for a brief critical review of the literature on this question.

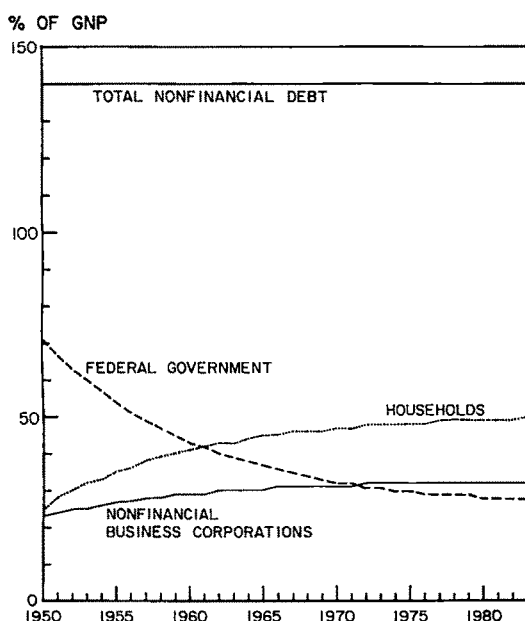


FIGURE 3. SIMULATED OUTSTANDING DEBT RATIOS FOR U.S. NONFINANCIAL BORROWERS

ing for the federal government, households, and nonfinancial corporations and with  $S_t = 1.40$ . Trends over the entire postwar period for  $S_g$ ,  $S_h$ , and  $S_c$  are then generated assuming a constant nominal GNP growth rate of 10 percent, a stable  $S_t$  at 1.40, and constant proportional shares of net new debt issues for the three sectors. The values used in setting each sector's constant proportional share of net debt issues are their actual mean proportions over the postwar period. As can be seen in Figure 3, the simulated outstanding debt ratios replicate the actual patterns quite closely: we observe both the inverse covariation between the ratios beginning in the initial postwar years and the dissipation of that pattern by the mid-1970's. Hence, this exercise affirms that what is reflected in the actual trends of the sectoral outstanding debt ratios is the grafting of relatively stable proportions of debt issuance onto the skewed outstanding debt structure of the immediate postwar years. The observed inverse covariation of the actual sectoral ratios does not therefore reflect any corresponding inverse relationship in the sectors' rates of debt issuance over the postwar trend.

### III. Debt Ratios and Financial Activity

#### A. The Impact of Government Deficits

Frequently the debt ratios are examined in order to provide an empirical framework for analyzing the impact of rising federal government deficits. Almost always when this is done, attention is focused only on the outstanding debt ratios. In Friedman's work (especially 1984), where the debt ratios are analyzed most fully in this regard, the central concern is the rising value of the government's outstanding debt ratio  $S_g$  that will occur over the next four years through even the most conservative estimates of forthcoming government deficit levels. The implications of this pattern for Friedman are derived from his analysis of the debt ratios' behavior throughout the postwar period.

He first of all believes that the long-term stability of the total outstanding debt ratio  $S_t$  represents a fundamental equilibrium point for the economy: the continuation of this fixed ceiling for  $S_t$  becomes the major assumption in his analysis. Based on this assumption, an increase in one sector's debt ratio becomes possible only through a declining ratio for some other sector(s)—hence the inverse covariation between the sectoral ratios over the postwar period. From this perspective then, the postwar trend decline of  $S_g$  served to release scarce financial resources from the federal government, and thus to permit the household and corporate rates of debt issuance (their "reliance on debt") to rise. In Friedman's view, we observe this increasing reliance on debt by households and corporations through the upward trends for  $S_h$  and  $S_c$ . The prospect of a rising  $S_g$  over the next four years thus implies that, in the absence of a fully accommodating (and inflationary) monetary policy, the financial resources available to households and corporations will become increasingly scarce and costly. This is particularly harmful in Friedman's view because corporations have relied upon debt increasingly since the mid-1960's to finance investment spending. He believes investment growth will be constrained with  $S_g$  rising, and as a result, advances in productivity generated through increasing the capital stock will be inhibited.

### B. *A Critique and Reformulation*

While this perspective is internally consistent, it is misleading in several important ways due to its exclusive focus on the outstanding debt ratios. The problem is not simply that debt-financing patterns are overlooked; equally distorting is that the conceptual distinction between debt financing and debt accumulation is not recognized. When the debt-financing patterns are explicitly incorporated into the analysis, a significantly different, and I believe more valid, perspective on federal deficits becomes possible, as does a more accurate assessment of U.S. financial conditions in general.

First, once we recognize the phenomenon of debt financing as distinct from debt accumulation, it becomes clear that the financing ratios, not the outstanding debt ratios, measure rates of debt issuance, the reliance on debt as a source of finance during any given time period. Of course, the outstanding debt ratios reflect the financing patterns to some extent. However, as I have argued above, the behavior of these ratios is also skewed by the effects of wartime finance and inflation. While these effects may indeed alter the behavior of borrowers and lenders as well as the future value of the new debt, certainly they should not bear upon how the current period issuance of debt is measured.

Given this, it follows that the observed inverse covariation of  $S_g$  with  $S_h$  and  $S_c$  does not signify a corresponding inverse covariation in each sector's reliance on debt finance. In fact, as the debt-financing ratios show, the three sectors experienced similar patterns of debt issuance throughout the postwar period: stability through the mid-1960's as  $f_g$ ,  $f_h$ , and  $f_c$  were all trendless; and increasing debt issuance relative to *GNP* thereafter. From this perspective then, we perceive no necessity for the government's rate of debt issuance to fall in order for households' and corporations' debt issuance to increase.

An alternative view as to the capacity of the nonfinancial sectors in the aggregate to undertake increases in debt financing also emerges from this perspective. More specifically, with the total debt-financing ratio  $f$ , having risen since the mid-1960's, it appears

—contrary to the impression suggested by  $S_i$ —that no apparent limit has been defined as to how much net new debt the nonfinancial sectors can issue relative to *GNP*. This reinforces the idea that increasing debt financing by the federal government does not prevent other sectors from acting similarly.

The magnitude of the federal government's postwar financial activities also assumes different dimensions once debt-financing ratios are taken into account. By any measure, the current rate of issuance by the federal government is unprecedented for the postwar period. But it is not the case—as is inferred by considering solely the trend for  $S_g$ —that this behavior signifies the reversal of a long-term pattern of decreasing debt reliance by the government. Rather, it actually represents the continuation, albeit at an accelerated rate, of a trend that began in the 1960's toward increasing government debt financing. By this measure, in other words, we observe that the government's presence in financial markets—and its corresponding impact on macroeconomic behavior through deficit spending—has been rising, not falling during the past fifteen years.

Finally, the relationship between the nonfinancial corporations' reliance on debt financing and their capacity to enlarge the capital stock is also clarified by examining the debt-financing data. It is true, as the pattern of  $f_c$  implies, that nonfinancial corporations have relied increasingly since the mid-1960's on debt financing to undertake investment spending. But this increase in debt reliance has not been associated with rising rates of real fixed-investment growth. Rather fixed-investment growth for nonfinancial corporations was significantly higher in the years prior to the mid-1960's than afterward (mean values for real annual fixed investment growth were 5.4 percent between the cyclical peak years 1949–66 and 4.3 percent for 1966–83). Increasing debt financing by the corporations, in other words, has been associated with a *declining* fixed-investment growth rate. It is therefore not apparent that high rates of corporate debt financing will bring increasing rates of capital formation, or that decreasing rates of debt financing will necessarily retard investment growth.

### C. *Further Considerations on Financial Instability*

Clearly, an alternative perspective of contemporary U.S. financial conditions emerges when the debt-financing patterns are taken into account. Moving beyond the single question of federal deficits, this perspective can serve as a basis for clarifying several other issues linked to the problem of financial instability. First, what is most evident in observing the debt-financing ratios is that a significant change on the demand side of credit markets—an upward shift in aggregate reliance on borrowed funds—did emerge after 1966. This development, unobservable through the outstanding debt ratios, is nevertheless fully consistent with other important changes in the financial structure beginning in the mid-1960's: on the supply side, the growing importance of financial innovation and liability management; and more broadly, the emergence of increasingly serious "credit crunches" and other crisis episodes (see Albert Wojnilower, 1980). These latter phenomena are much less comprehensible if one begins with the conception that aggregate reliance on debt has been essentially stable over the past thirty-five years.

By recognizing the debt-financing and outstanding debt ratios as analytically distinct, we may also now distinguish two separate forces affecting the stability of the financial structure on the demand side: changes in the reliance on debt, and changes in the *debt burden*. While the debt-financing ratios are superior indicators of debt reliance, clearly the debt burden must be measured on the basis of the accumulated stock of debt.<sup>3</sup> Establishing this distinction in turn brings into focus one fundamental aspect of inflation's impact on the financial structure, an effect which arises through inflation's asymmetric impact on the stock and flow ratios. Because of this asymmetry, the nonfinancial sectors' increasing reliance on debt over the past twenty years has not engendered similar increases in their debt burdens. Thus, how-

ever else inflation may affect financial conditions, it exerts a stabilizing influence insofar as it reduces net borrowers' real debt burden. Similarly, disinflation (to say nothing of deflation) will exacerbate instability by increasing the real debt burden. These factors become particularly important if the nonfinancial sectors' reliance on debt continues to rise in the future while the inflation rate remains at a relatively low level. It follows also that to counter tendencies toward overindebtedness through means other than inflation will probably require that the nonfinancial sectors' debt-financing ratios be stabilized at a lower level than in recent years.

To explain the debt-financing ratios' upward trend thus emerges as a central problem. While I cannot even begin a systematic discussion of this issue here (I do attempt such with respect to nonfinancial corporations in my earlier paper), one important aspect of it can be highlighted as a point for further consideration. The data presented here show that the rise of  $f_t$  and its sectoral components has not occurred in association with increases in the growth rates for real investment or output, but rather with declining real growth for both. It therefore appears insufficient to explain the rise of debt financing through a standard neoclassical framework emphasizing price effects, specifically the declining real borrowing costs and/or debt burden which occurred through the 1970's. If these were the prime motivating forces, we would then expect the growth of real expenditure and output to rise along with the increase in debt financing. For similar reasons, it also appears inadequate to frame the problem in the manner of Hyman Minsky (1982) and Wojnilower, as an expression of a persistent boom psychology. Rather it seems that a more fruitful place to focus the analysis is at precisely the point stressed by Gurley and Shaw—the relationship between debt financing and income growth. Working from there, the increasing debt reliance would appear to have resulted primarily because of the declining capacity of real income growth—and the internal funds derived from income—to meet perceived expenditure needs.

<sup>3</sup> However, debt stocks themselves do not fully measure debt burdens, since interest payment obligations must be included as part of the burden.

## REFERENCES

- Friedman, Benjamin, M., "Managing the U.S. Government Deficit in the 1980s," in M. L. Wachter and S. M. Wachter, eds., *Removing Obstacles to Economic Growth*, Philadelphia: University of Pennsylvania Press, 1984.
- \_\_\_\_\_, "Debt and Economic Activity in the United States," in his *The Changing Roles of Debt and Equity in Financing U.S. Capital Formation*, Chicago: University of Chicago Press, 1982, ch. 6.
- Gurley, John G., and Shaw, Edward S., "The Growth of Debt and Money in the United States, 1800-1950: A Suggested Interpretation," *Review of Economics and Statistics*, August 1957, 39, 250-62.
- Minsky, Hyman P., *Can "It" Happen Again?: Essays on Instability and Finance*, Armonk: M. E. Sharpe, 1982.
- Pollin, Robert, "Alternative Perspectives on the Rise of Corporate Debt Dependency: The U.S. Postwar Experience," mimeo., September, 1984.
- Wojnilower, Albert M., "The Central Role of Credit Crunches in Recent Financial History," *Brookings Papers on Economic Activity*, 2:1980, 277-326.

# Private Credit Demand, Supply, and Crunches— How Different are the 1980's?

By ALBERT M. WOJNOWER\*

In the fall of 1980, I had the unusual privilege for a business economist of preparing a Brookings Paper on Economic Activity. In post-1950 U.S. business expansions and especially near business cycle peaks, I argued, aggregate private credit demand had been essentially interest rate inelastic and far in excess of the supply. As a result, downturns in credit use and general business activity had developed only after some blockage in the supply of credit—a sudden and unanticipated intensification of nonprice rationing, commonly labeled a “credit crunch.”

These often frightening crunches were triggered either by 1) credit or interest rate controls that sharply constricted the lending incentives or capacities of financial institutions, or 2) unanticipated default crises that had the same paralytic impact. Each such episode, however, prompted responses by the affected parties (borrowers, lenders, and regulators) to remove the “bottleneck” that had provoked the crisis. The consequent financial innovation, deregulation, and attendant broadening of the Federal Reserve's lender-of-last-resort functions greatly enlarged the supply of credit, intensified the inflationary potential of business upswings, and heightened the risk of a general financial collapse.

The plan here is to sketch the relevant experience subsequent to October 1980 when my earlier paper was presented. As with econometric models, the postsample observations do not fit as neatly as did the in-sample data. Nevertheless, I believe, the fundamental conclusions have stood up well.

## I. Credit Demand during Expansions

For those not accustomed to thinking of aggregate credit demand as innately huge

and inelastic, the stubborn persistence of rapid credit growth and business activity in the face of the 1979–80 financial turmoil, as well as their virtually instant revival following the classic crunch induced by the short-lived credit controls of 1980, was a great surprise. The controls were imposed on March 14; on July 3, the last remaining restrictions were abolished. During the spring, there was a sharp economic recession and drop in interest rates, but both ended in July. By the close of 1980, interest rate levels had already surpassed the new records they had established early in the year. From December 1980 through October 1981, notwithstanding a dramatic and persistent reduction in inflation, the bank prime loan rate hovered almost uninterruptedly in the 18–20 percent range.

But credit demand was not deterred. Business loan expansion, much of it to support sizable increases in real fixed investment, achieved and maintained record growth rates into the opening months of 1982, far beyond the onset of recession in midsummer 1981. The growth in loans slowed appreciably only when default calamities began to erupt in the spring of 1982. Consumer borrowing growth also accelerated dramatically, although it failed to regain previous peak rates. Some retardation in growth began during the spring of 1981, but no material setback occurred until late 1981, well after the business cycle crest in July.

Only the behavior of mortgage credit departed radically from earlier experience. In previous business cycles, mortgage credit growth had accelerated well along into the general business upswing, but in the 1980–81 recovery it began to diminish within a few months. This weakness, however, did not reflect any substantial decline or rate elasticity in mortgage demand, but rather some crippling constraints on the lenders (described in Section II below). In the 1983–84 upswing, by which time the blockages in

\*Managing director and chief economist, The First Boston Corporation, Park Avenue Plaza, New York, NY 10055.



mortgage supply had been removed, mortgage generation recovered strongly and remained vigorous, notwithstanding a level of effective interest rates that, in view of reduced general and house-price inflation, was probably higher yet than in 1980–81.

The business upturn that began in late 1982 provides a remarkable demonstration of how strong credit demand can be even in depressed circumstances. The initial thrust of the business recovery clearly depended on surging demand for homes and cars, two notoriously credit-dependent sectors. An enormous burst of private credit growth began at a time when consumer and producer prices had remained essentially unchanged for some six months, and while mortgage, consumer, and business loan interest rates—and unemployment—all exceeded 10 percent!

In 1983 and 1984, the expansion of private (and federal) credit continued apace, although nominal interest rates generally rose and inflation expectations (albeit not the measured rate of inflation) allegedly abated. No doubt there exist interest rate levels high enough to curb private credit demand, but the experience of the 1980's to date, featuring good times and bad, and intervals of double-digit as well as zero inflation, suggests they lie well above the range of recent observation.

The proliferation of floating-rate lending has further attenuated whatever rationing power interest rates may exert. For at least several years now, some 75 percent or more of large over-one-year bank loans have been made at variable rate. But lately the major part of corporate bond issuance also has become variable-rate and/or short-term in character. Fixed-rate issues with over ten years to maturity are a vanishing breed, comprising only 21 percent of nonconvertible corporate bond issuance in 1984, down from about 35 percent in 1981–83, and 60 percent or more in earlier years.<sup>1</sup> In my earlier paper

I noted the apparently indestructible “optimism” of industrial executives and institutional investment managers that interest rates will fall. Perhaps lenders now see more need for protection against possible interest rate increases, but business borrowers apparently remain confident that rates will decline.

The home mortgage market is another sector in which floating rates have just recently come to predominance. In recent months, some two-thirds of new home mortgages have carried variable rates. But in mortgages, this seems to be mainly at the lender's rather than the borrower's initiative. Consumers must be “bribed” to accept variable rates. The essential “sweeteners” include a lower “base” rate, sizable initial-year discounts, caps on rate increases and/or levels, provisions for “negative amortization” to hold down monthly payments, and so on. This accords with my 1980 view that the borrowing behavior of households, though not particularly interest rate sensitive, is much more restrained than that of business firms and financiers.

The great majority of variable rate loans of all kinds is indexed to a one-year or shorter-term interest rate. What role remains for interest rates in rationing credit demand when most credit is either short-term, or indexed to a short-term interest rate? Do long rates (and the yield curve) still matter if hardly any private borrowing takes place at these rates? Do even short rates matter much? Short-term changes in short rates should affect only short-horizon decisions.

In actuality, interest rate fluctuations now appear to exert their principal impact by altering the cash positions of floating-rate borrowers. Mortgage debtors, however, are cushioned against rate increases by indexing lags, caps, and negative amortization. Among business debtors, rises in rates seem merely to strengthen the consensus that rates must come down soon. Thus firms always are inclined to hold on just a little longer rather than to retrench. The rising rates simply add an involuntary (by definition, interest-inelastic) component to credit demand to finance the unexpectedly higher interest payments.

The merry-go-round spins until insolvency looms. To become effective, monetary re-

<sup>1</sup>Based on the first ten months of 1984. Figures by courtesy of Dr. Henry Kaufman, Executive Director, Salomon Brothers, Inc.

straint must threaten widespread defaults. Only lately is it becoming generally recognized that the floating-rate mode converts interest rate risk to credit risk. (See my earlier paper, p. 296. For a formal treatment, see Anthony Santomero, 1983.)

## II. The Supply Side

An important reason why lenders are eager to innovate is their awareness of the chronic excess demand for credit. Lenders wish to be deregulated so as to be free of constraints that prevent them from lending more. During 1980–82, new deregulation was relatively slow, but dramatic further steps, particularly the introduction of money market deposit and Super-NOW accounts with no interest rate restrictions, took place in late 1982 and early 1983. The credit boom of 1983–84 is one consequence.

Indeed, from the standpoint of the individual depository institution, the recent abolition of most interest rate controls on consumer deposits has made more rapid expansion a matter of necessity as well as choice. An earnings squeeze has been created from which “growing your way out by lending more” seems the only hope of escape. But of course the intensified competition for loans narrows profit margins further, and this at a time when hostile capital markets leave many banks and thrifts more than ever dependent on internal funds for strengthening their capital. So long as interest rate ceilings held deposit rates at below-market levels, depository institutions had an easier way out. The cushion provided by a reliable stock of cheap deposit “raw material” gave them the option of building up earnings and capital by means of a “play safe” policy of investing only in relatively low-yield safe loans and government securities. But in the new environment, safety is a luxury most lenders cannot afford.

In the large, of course, the total quantity of credit remains in principle determined (intentionally or not) by Federal Reserve policy. But the Federal Reserve is a rather complex “endogenous” variable. Such is especially the case whenever, as has been the predominant view for at least twenty-five

years, interest rates are regarded as high. When rates are thought to be high, or when the financial system seems threatened by defaults, the Federal Reserve will be relatively timid in tightening. If, concurrently, lenders have reason to become more aggressive, they in effect lead the authorities to what by hindsight becomes recognized as overly expansionary policy.

As the last recession drew to a close in late 1982, it was widely conceded that the financial structure was in debilitated condition. The banking system, thrift institutions, parts of the securities industry, and many businesses had recently brushed disaster. Not only should the demand for credit have been weak, but also the supply. Domestic borrowers should have experienced credit access problems similar to those of the *LDCs*. Yet by mid-1983, the growth rate of domestic private debt resumed its prerecession pace. By year end, it was growing at a rate previously reserved largely for the “elderly,” inflation-ridden stages of business expansion.

Dramatic changes occurred in the character of many credit markets. There was a prodigious outburst of bank lending to finance mergers and acquisitions, particularly so-called “leveraged buyouts.” Other types of business loans also soared. In the first half of 1984, net debt of nonfinancial corporations enlarged at a \$181 billion annual rate; the previous peak year (1981) was \$103 billion. This does not include the burgeoning of credit guarantees issued by banks, for which there are no statistics.

Explosive growth also took place in consumer credit. The dollar amount of the increase for the first half of 1984 (seasonally adjusted) exceeded any previous full year's gain. In the face of significant increases in the general level of interest rates during 1983 and the first half of 1984, the intense lender competition actually forced installment credit charges down; as of mid-1984, they were still near their post-1980 lows.

It was, however, the single-family home mortgage market that was most radically transformed. In 1980–82, as indicated before, both the lending incentive and capacity of thrift institutions had been crippled. Because short rates were chronically and atypi-

cally higher than long rates, using short-term funds to finance mortgage lending was unprofitable. Also, thrift deposit rates were still regulated in such a way that they could not be relied upon to remain, and often in fact were not, fully competitive with alternatives such as money market funds. Furthermore, although mortgage rate ceilings set by state usury laws had been preempted by federal legislation in March 1980, the states had been given three years to reinstitute rate ceilings if they chose. Thus it was uncertain whether the provisions of high rate or floating-rate mortgages would stand up. At many institutions, the mortgage rates posted were those at which they *might* lend to preferred clients, rather than those at which they invited business.

All this had passed by the beginning of 1983. Short rates were below long rates, and the relevant restrictions on deposit rates had been or shortly were to be abolished. Legal limits on mortgage rates were effectively defunct. Many would-be borrowers might have been unable, however, to meet income tests at the prevailing rates. Lenders responded resourcefully. Partly reflecting a legal mandate that suddenly became commercially useful, they became more enthusiastic about counting second incomes (usually wives'). Also income tests were eased, directly as well as indirectly. Substantial rate discounts are widely offered for the first year or two, which help bring down the initial monthly payments into qualifying range. In a dramatic turnabout, the growth rate of home mortgage debt surged from a near-record low of 4 percent in the second half of 1982 to a 12 percent rate of gain only a year later, a pace that was still being maintained in the third quarter of 1984, almost two years into the business upswing (see Figure 1).

The impetus toward more aggressive lending is by no means confined to the large institutions. Under deregulation, new or unknown banks may and do offer higher deposit interest rates than well-known banks. The unknown bank may have localized high-return outlets not accessible to larger lenders; or it may be willing to take greater credit risk or accept lower markups; or it may have (or thinks it has) lower costs. The

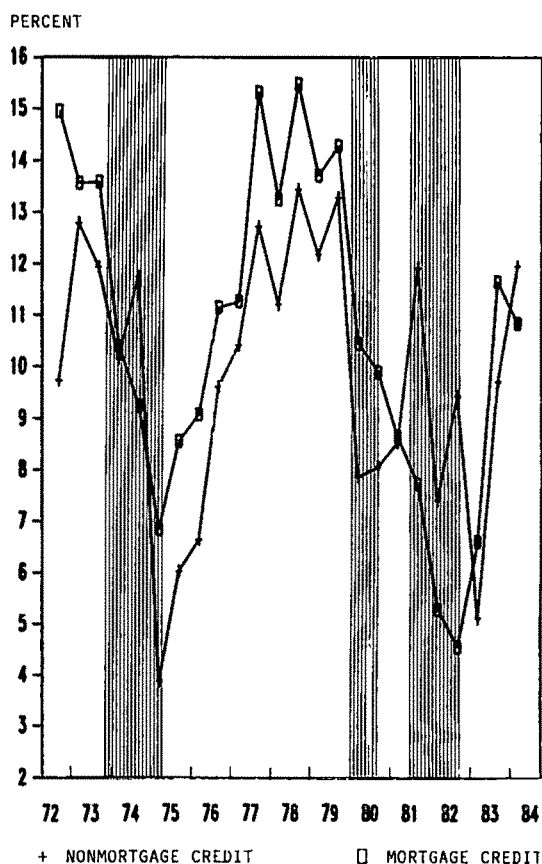


FIGURE 1. GROWTH RATES OF PRIVATE NON-MORTGAGE CREDIT AND PRIVATE MORTGAGE CREDIT (Semiannually seasonally adjusted annual rates. Shaded areas are business recessions.)

result is the more intensive exploitation of lending opportunities as well as the narrowing of margins and easing of credit standards (nonprice terms).

Conditions in the credit industry have many instructive analogies with the newly deregulated airline business. One is that new "fly-by-night" airlines can prosper because their FAA certification is equivalent to that of established lines; fly-by-night banks can prosper on the strength of their official deposit insurance. At a price, many people probably would be willing to fly uncertificated planes and to deposit funds in uninsured banks; but the general public (wisely) is unwilling to tolerate the social consequences of plane or bank crashes.

For an individual airline, the competitive response to deregulation, short of going out of business, is to price so as to fly fuller planes more continuously. That, subject to technological and environmental constraints, is presumably a good thing. But do we wish financial institutions always to be loaned to the hilt? A lesser rather than a higher rate of credit expansion is implied by the national goal of price stability.

### III. Crunches and Crises

During the early 1980's, I conclude, the credit markets have continued to be characterized by an explosive mixture of chronic excess demand and rapidly expanding supply. Yet were that the whole story, only a credit crunch—a major interruption in credit supply—could have appreciably retarded private credit expansion and triggered a recession. Unquestionably, however, the business downturn that set in during the summer of 1981 was unheralded by any credit crunch. Although the later succession of credit crises that developed in May-August 1982, summarized below, may warrant the crunch label, by then the recession was already long under way.

How could this have happened? As already indicated, the explanation probably lies mainly in the anomalous behavior of mortgage credit in 1980-81. There was a crunch in mortgage credit, but not in other types of credit. As a result, the recovery in homebuilding began to be aborted almost immediately and neither aggregate credit expansion nor the economy ever regained full momentum. If subsequent mortgage-market innovation and deregulation have permanently enhanced the viability of mortgage generation under financial adversity, as I believe they have, the 1981 "quiet" business cycle crest without a general credit crunch will not become typical.

In the absence of credit supply squeezes due to regulatory constraints, lender "growthmanship" has, as was to be expected, multiplied the "accidents" that activated lender-of-last-resort support. While some crises of the earlier 1970's (such as the Penn Central, Franklin National Bank, and Hers-

tatt debacles), provoked credit crunches and recessions, reaction to subsequent major default problems demonstrated that, by 1980 if not sooner, the public had come to take for granted that the authorities would not allow any sizable financial actor to default. As a result, although several recent shocks have carried ample potential for catastrophe, their adverse macroeconomic impact has been mild.

One such "accident" was the May 1982 collapse of Drysdale & Co., a minor government securities dealer. Although the firm was small, the amounts and ripple effects were large. One of several dangerous consequences was to cloud the lender's claim on government security collateral used in repurchase agreements, the principal form of credit used by dealers to finance their positions. Had not the Federal Reserve intervened with alacrity, the functioning of the government securities market would have been seriously impaired.

Soon after came the Penn Square disaster. This was the failure of a smallish Oklahoma bank that had managed to sell billions of questionable loans to major banks and was also a depository for a number of other small banks. The regulators attempted to limit their involvement by refusing to accept liability for deposits exceeding \$100,000, nor did they provide financial support for a merger of the bank. For the first time since the creation of the FDIC, large depositors in an insured bank actually suffered losses. Whether this helped to hasten or to forestall later failures of other depositories remains a debatable issue. In any event, when the authorities later were tested by runs on truly giant institutions, namely the Continental Illinois Bank and the savings and loan subsidiary of Financial Corporation of America, they really had no choice but to confirm that *all* deposits in institutions of national significance are *de facto* insured. For Continental Illinois, in addition, the regulators felt obliged also to guarantee, over the public objection of Treasury Secretary Regan, the credit market liabilities of the holding company that owned the bank. Although it was carefully stipulated that no precedent was intended, this allowed another large camel to

slip its nose under the tent. The recent entry of major insurance companies into the business of insuring banks and bond investors against loan defaults represents another effort to stretch the safety net. Now, it can be presumed, the authorities will have to interdict a cascading of defaults if only to save the insurance industry.

Also in 1982, major problems arose in connection with the rescheduling of the debts of Mexico, Brazil, and other foreign countries, default on which would have decimated the capitalization of leading American banks. The Mexican crisis came to a head in the summer, when the U.S. financial system was already in a state of jitters. The Treasury and the Federal Reserve promptly undertook a rescue operation that reassured all concerned not only of the safety of their deposits, but also that the affected banks would not have to abruptly curtail their operations. As a result, the many subsequent foreign debt "crises" have had little if any general market impact.

The cumulative effect of the responses to these and other calamities has been to make the authorities the implicit guarantors of nearly the entire spectrum of widely held debt obligations. This may not be of great concern while inflation remains relatively subdued. Eventually, however, a more pronounced collision is likely between these protective functions and anti-inflation goals. It is difficult to inhibit lending by financial institutions and markets while providing them unconditional shelter.

Mainly as a consequence of the international debt problems, Congress in late 1983 mandated more stringent bank capital regulation. The regulators' response, however reluctant, of closer supervisory scrutiny, public criticism of certain types of merger and

mortgage lending, and promulgation of more formal, and for large banks stiffer capital requirements, appears to be exerting some restraining impact on the growth of credit. This approach implies renewed experimentation with credit-control (crunch-producing) techniques whenever the next period of sharply restrictive monetary policy eventually reaches the pitch at which it threatens business debtors with insolvency and home mortgage debtors with higher monthly payments. At such a time, the limitation of credit growth by means of binding capital-to-asset ratios would skyrocket interest rates. In addition, lenders would further restructure their portfolios so as to minimize holdings of governmental securities, mortgages, and other low-risk, low-return credit. The likely political reaction would be toward more specific regulation of interest rates and portfolio composition.

I have long argued that deregulation of financial markets, particularly as respects deposit interest rates, will ultimately spawn new and broader control machinery more onerous than its forebearers. That federal authorities with an ideological commitment to full financial deregulation find themselves newly immersed in intense "hands-on" supervision is strong evidence that such a trend to reregulation has started.

## REFERENCES

- Santomero, Anthony M., "Fixed Versus Variable Rate Loans," *Journal of Finance*, December 1983, 38, 1363-79.
- Wojnilower, Albert M., "The Central Role of Credit Crunches in Recent Financial History," *Brookings Papers on Economic Activity*, 2:1980, 277-326.

## MACROECONOMIC ANALYSIS OF LEADING INTERWAR AUTHORITIES†

### Marriner S. Eccles, Chairman of the Federal Reserve Board

By L. DWIGHT ISRAELSEN\*

The inherent instability of capitalism may be corrected by conscious and deliberate use of three compensatory instruments, taxation, varying governmental expenditures, and monetary control.  
*Marriner S. Eccles, 1935*

Yours was the only revolution on record that entered government by way of a central bank.

*John Kenneth Galbraith to  
Eccles, 1976*

On July 29, 1983, the Federal Reserve Building in Washington, D.C. was formally named in honor of Marriner S. Eccles, who served as Governor of the Fed, 1934–36, as Chairman of the Board of Governors, 1936–48, and as member of the Board, 1948–51. While Eccles was honored primarily for his struggle to maintain Federal Reserve independence to conduct monetary policy, his role in introducing “compensatory” monetary and fiscal policies—modern macroeconomic stabilization policies—was undoubtedly of equal importance. In tracing the development of Eccles’ macroeconomic philosophy, this study identifies him as one

of the earliest American precursors to Keynes, and as the most important figure in the introduction of “Keynesian” economic policies in the United States.

#### I. Changing Views

We must acknowledge that progress comes only through toil, economy and thrift, and that these alone are the motive power which creates the enduring structure.  
*Eccles, 1925*

The matter of economy is negative, the matter of spending is positive, and we have been doing the negative thing rather than the positive. We have been preaching the negative doctrine.... Our depression was not brought about as a result of extravagance.... The difficulty is that we were not sufficiently extravagant as a nation. We did not consume what we were able to produce.  
*Eccles, 1932*

In February, 1933, the Finance Committee of the United States heard testimony on the causes and cures of the depression. While farmers argued for remonetization of silver as a means of recovery, and labor spokesmen suggested reduced hours and work weeks, the majority of the 46 prominent Americans who testified were of the opinion that the depression represented the workings of natural economic law, a punishment for the “extravagance” of the 1920’s, and that to interfere with the cycle of boom and bust was to invite disaster. Balancing the federal budget in order to “restore confidence” was seen as the only prudent course open to government, a policy reflected in the 1932 political campaign, in which both major parties advocated a balanced budget as the key to economic recovery. In an effort to reduce the deficit,

†*Discussants:* Raymond W. Goldsmith, Yale University; Takafusa Nakamura, University of Tokyo.

\*Associate Professor of Economics, Utah State University, Logan, UT 84321. I am indebted to personnel at the Special Collections Library, University of Utah, for assistance in working with Eccles’ papers, the Marriner S. Eccles Collection (hereafter *MSE*). I also benefited greatly from conversation and correspondence with economists acquainted with Eccles and his work, particularly Lauchlin Currie, Evsey D. Domar, Milton Friedman, John Kenneth Galbraith, Charles P. Kindleberger, Richard A. Musgrave, and Herbert Stein, none of whom should be held responsible for my inferences and conclusions.

Congress had passed in 1932 what was to that point the largest tax increase in U.S. history. Toward the end of the hearing, Eccles was called to testify. Eccles, 43-years-old, was a successful Utah banker and industrialist, a "conservative Republican" whose formal schooling had ended after three years of high school, but whose business acumen had enabled him to bring his commercial and banking interests through the worst years of the depression relatively unscathed.

In his testimony, Eccles identified the cause of the depression as an insufficiency of effective demand, rather than punishment for past extravagances, loss of confidence, or workings of natural law. The cure, Eccles stated, was a restoration of sufficient spending to purchase the quantity of goods which it was possible to produce at full employment. Because the profit motive could be expected to lead individuals, business, and financial institutions to make decisions which would further reduce spending, hence income and employment, the government, motivated not by profits, but by the welfare of the public, must compensate by spending more. "I see no way of correcting this situation except through Government action," Eccles declared (*MSE*, Senate Finance Committee, *Investigations of Economic Problems*, "Statement of M. S. Eccles, President First Security Corporation, Ogden, Utah," February 24, 1933, p. 712). He then proceeded to outline a five-point program of unemployment relief, public works, agricultural allotment, farm mortgage refinancing, and permanent settlement of interallied debts to deal with the immediate problems of the depression. He also proposed a plan for long-run economic stability that included unification of the banking system under the Federal Reserve and the creation of an agency to guarantee bank deposits; tax reform to achieve a more equitable distribution of wealth and purchasing power; passage of national child labor, minimum wage, unemployment insurance, and old-age pension laws; federal agencies to approve all new capital issues offered to the public and all foreign financing, all means of transportation, and all means of communication to insure their operation in the public interest; and a national planning board to coordinate public and private economic ac-

tivities (see also pp. 712-33). Eccles' testimony was received by the Finance Committee with a mixture of interest, skepticism, disbelief, and outright hostility. Three years later, a cover story in *Time Magazine* evaluated his 1933 proposals in the following terms: "Eccles laid before a Senate committee a plan, which turned out to be nothing less than a detailed blueprint of the New Deal. Only one Eccles suggestion has not materialized—official cancellation of War Debts" (February 10, 1936, p. 60).

Marriner Stoddard Eccles was the eldest son of the second wife of David Eccles, Utah's first native millionaire. David Eccles, who had been illiterate when he emigrated from Scotland to Utah at the age of fourteen, left an estate appraised at more than \$7 million when he died in 1912 (see Leonard Arrington, 1975). Marriner Eccles' success in consolidating and managing his father's estate has been documented in Eccles' autobiography, edited by Sidney Hyman (1951).

In addition to wealth, Eccles inherited from his father a set of beliefs about the proper roles of individuals and government in the functioning of the economic system. In the automatically functioning, self-adjusting, capitalist economy, the role of government should be limited to "maintaining confidence" through strict budget-balancing, while the greatest benefit would be received by those individuals who worked hard, practiced strict economy, and invested prudently (Hyman, 1951, pp. 4-5, 21, 37, 51). "We must acknowledge," Eccles told Utah bankers in 1925, "that progress comes only through toil, economy and thrift, and that these alone are the motive power which creates the enduring structure" (*MSE*, written address, June 1925, p. 3). As the crash of 1929 deepened into prolonged depression, Eccles was forced to reevaluate his thinking. Early in 1931, he recalls, "I saw for the first time that though I'd been active in the world of finance and production for seventeen years and knew its techniques, I knew less than nothing about its economic and social effects. The discovery of my ignorance, however, did not by itself lead anywhere.... As an individual I felt myself helpless to do anything" (Hyman 1951, pp. 54-55). Eccles had run aground on the shoals of macroeconomics.

The reformulation of his thinking was reflected in his public addresses. In a speech to bankers in the spring of 1931, Eccles said, "The modern system by which society supplies its wants...is a wonderfully effective organization when in balance..., but if anything happens to throw it out of balance it is possible to have millions of people unable to buy the products of others because they cannot sell their own" (*MSE*, written address, March 26, 1931, pp. 5-6). Eccles had by this point rejected the idea of the automatic restoration of economic prosperity through the workings of the invisible hand of narrow self-interest. He had discovered the fallacy of composition, and had concluded that "intelligent and courageous" open-market purchases by the Fed could have averted the drastic deflation which followed the crash. He had also concluded that underconsumption, not overproduction, was the basic cause of the depression (see also pp. 4-5, 7-8).

During the period 1931-33, Eccles developed and expanded the underconsumption theory, and by 1933 had arrived at the essential framework of Keynesian analysis and policy. The evolution in his thinking can be identified in his public addresses of 1932 and 1933. In a 1932 speech, Eccles declared, "Our depression was not brought about as a result of extravagance.... The difficulty is that we were not sufficiently extravagant as a nation. We did not consume what we were able to produce" (*MSE*, written address, June 17, 1932, pp. 2-3). Eccles mentioned the fallacy of composition problem and the futility and perversity of government efforts to balance the budget, which could only lead to further unemployment. "Just to the extent that unemployment increases," said Eccles, "just to that extent are you going to find it more impossible to...balance any budget" (p. 4). Popular theories of causes of the depression were dismissed: "These are not acts of God, they are mistakes of man" (p. 6). Traditional theories of economic recovery were also discarded. "The theory of hard work and thrift as a means of pulling us out is unsound economically. True hard work means more production, but thrift and economy mean less consumption.... Now for the solution to our problem," said Eccles, "How are you going to put these people back to

work? There is only one agency in my opinion that can turn the cycle upward and that is the Government. ...[T]he Government, if it is worthy of the support...of its citizens, must so regulate, through its power of taxation, through its power over the control of money and credit,...the economic structure so as to give men...the opportunity to work" (see pp. 5-6).

Eccles had further refined his ideas by 1933. On economic recovery through spontaneous revival of investment, he said, "The assumption of spontaneous revival through new investment has always rested on the fallacious belief that people and banks will not indefinitely hold money in idleness" (*MSE*, written address, October 27, 1933, p. 3). On the notion that a shortage of currency in circulation was prolonging the depression, Eccles stated, "There is no shortage of currency in circulation.... The need is not for more money, but for more spending" (see pp. 5, 2). To the view that recovery was dependent on the establishment of "sound money," Eccles commented,

For the past two years or more we have had the painfully sound dollar measured by its purchasing power in terms of goods and services. The sounder it got the further prices fell and the more unemployment increased. Had the policy of economy and budget balancing on the part of the Government continued, it would have soon been so sound that all of our credit institutions would have been closed, there would have been no bank money and all of the people would have been starving to death with an abundance of everything for everybody, or at least the willingness and power to produce it.

[See pp. 10-11]

Self-interest cannot be relied upon to create recovery, Eccles told the Finance Committee, since "if we leave our 'rugged individual' to follow his own interest under these conditions he does precisely the wrong thing" (*MSE*, Senate Finance Committee, *Investigations*, February 24, 1933, p. 719). The decline in spending and investment since 1929 "could have been prevented by action of the Government which is the only agency which could continue spending money without re-



gard for profit.... Financial fuel is piled up—The Government, and not the bankers, must apply the torch. Motives of public welfare must lead us out of the present depression as greed and war have led the world out of past depressions,” Eccles said (*MSE*, written address, October 27, 1933, pp. 2, 4).

## II. Compensatory Economics

The government must be looked upon as a compensatory agency in this economy to do just the opposite of what private business and individuals do. The latter are necessarily motivated by the desire for profit. The former must be motivated by social obligation.

*Eccles, 1936*

By the mid-1930's, Eccles' compensatory policy recommendations were based on a sophisticated macroeconomic analysis which covered the consumption function; the multiplier; a distinction between the relative sizes of the government expenditure and transfer-tax multipliers; leakages and injections; causes of inflation; liquidity trap; velocity; the transmission mechanism of monetary influences; the Phillips curve relationship; the relationships among wage increases, productivity increases, and inflation; the role of inflationary and deflationary expectations; income and wealth distribution effects; the coordination of monetary, fiscal, and incomes policies; and the interrelationships between domestic stabilization policies and international movements of goods and capital. While he was not interested in the construction of a formal model, all of the elements of Keynesian analysis, with the possible exception of the accelerator, may be found in his speeches, letters, and memos.

All policies which came under Eccles' scrutiny were examined for stabilization implications. As an example, Eccles felt that Social Security taxation should be deliberately controlled in a countercyclical fashion, with increases in rates during booms and reductions during depressions. Taxation in general should be used mainly as a means of redistributing income from wealthy individuals and corporations to low- and middle-class

consumers who had, Eccles believed, higher marginal propensities to consume. He said in 1933,

The fundamental economic plans, when they are finally established, will of necessity center in the distribution of purchasing power and in the allocation of income between investment and expenditure.... They will involve relief of taxation that rests on the consumer... [and] the establishment of heavy income taxes especially in upper brackets. They will involve heavy taxation of undistributed corporate surplus, to force corporate income into dividends and taxes. [p. 7]

A good summary statement on compensatory policy was delivered by Eccles in 1935, when he declared his hope that “the inherent instability of capitalism may be corrected by conscious and deliberate use of three compensatory instruments, taxation, varying governmental expenditures, and monetary control.... It should be evident by now,” he said, “that simple maxims and rules of thumb are not sufficient” (*MSE*, written address, February 16, 1935, pp. 21–22).

Eccles saw in the instability of capitalism the seeds of destruction; in compensatory policy he saw the mechanism of salvation.

If we regard capitalism simply as a particular economic organization of society, our defense of, or attack on, that organization must be directed toward its effectiveness—its ability to satisfy in an adequate and equitable fashion the material needs of mankind. If it cannot be defended on these grounds it is doomed.... Private enterprise today is on trial solely because it is not producing the goods it has the capacity to produce and because it is not providing a more equitable distribution of the goods it is producing. [pp. 2–3]

The major threat to capitalism, Eccles believed, lay in the creation of a large group of unemployed. “These people no longer have any stake in preserving our present economy. They have nothing to lose. And if this condi-

tion persists...neither you nor I will have anything to lose" (pp. 2-3). "You have got to take care of the unemployed," he told the Finance Committee, "or you are going to have a revolution in this country" (*MSE*, Finance Committee, February 24, 1933, p. 733).

### III. Influences

My own viewpoint has sometimes been erroneously identified with that of Mr. Keynes, doubtless to his embarrassment. *Eccles, 1939*

I know of no professors whose writings have influenced me. *Eccles, 1949*

Eccles' contention that he arrived at his economic philosophy without having read Keynes (see Hyman, 1951, pp. 131-32) is accepted by his biographers (see Hyman, 1976, p. 128; Dean May, 1981, pp. 53, 58-59; Herbert Stein, 1969, p. 148). He had, in fact, read *something* by Keynes, as he quoted Keynes in 1933 on the difficulty of gaining public acceptance of deficit spending except in wartime (*MSE*, written address, October 27, 1933, p. 8). By this time, however, Eccles' ideas were already well-formulated, and he was actively searching for evidence and confirmation, such as he had found in the writings of Foster and Catchings (Hyman, 1976, pp. 93-94). With this minor exception, Eccles had apparently read nothing by Keynes before he came to Washington as Assistant to Treasury Secretary Morgenthau in 1934. His later exposure to Keynes' works was also limited. As Lauchlin Currie, Eccles' first and most important economic advisor recalls, "[Eccles] never read Keynes' *General Theory*, and he accepted the Keynesian arguments, when I summarized them [in a written review prepared for Eccles in November 1936], as a matter of course. Nothing new!" (Currie to author, letter, August 24, 1983). Eccles claimed to be innocent of any academic influences. "I know of no professors whose writings have influenced me," he stated in 1949 (*MSE*, Eccles to William Merrill, letter, September 22, 1949). The use of the word "professors" is instructive, as Eccles went on to identify eight individuals with whom he

worked at the Fed who might "have possibly had some influence on my thinking." Among those were some, Lauchlin Currie, Alvin Hansen, and Richard Musgrave, who had been and/or would be "professors." Musgrave contends that "economists undoubtedly influenced [Eccles'] thinking,..." His discussion of public debt in particular strongly reflects Hansen's position" (Musgrave to author, letter, September 22, 1983). These influences apparently came through the working relationship, rather than through "academic" writings. Evsey Domar recalls that Eccles never came to the Federal Reserve Seminar, which was established during the war with Keynes being the first speaker. "Somehow," Domar writes, "I have the impression that he did not care much for academic economists" (Domar to author, letter, July 6, 1983). Likewise, states Milton Friedman, "the academic world was not influenced by Eccles" (Friedman to author, letter, July 22, 1983). The academic world, however, was very much aware of Eccles' analysis and policy, as is indicated by his correspondence with many prominent economists. Irving Fisher, then Professor Emeritus at Yale, was one with whom Eccles had correspondence on several occasions. Although Fisher had written a very favorable three-part essay on Eccles' views in 1935 (*MSE*, Irving Fisher, "The Mind of Mr. Eccles," 1935, ms.), Eccles did not always agree with Fisher's views. In 1938, for example, Eccles, in a draft reply for Franklin Delano Roosevelt's signature, pointed out the flaws in a plan strongly urged on the president by Fisher to bring about economic recovery by monetizing the float. In submitting the draft, Eccles commented on such plans pressed on the president, and concluded, "I am returning also, with a suggested reply as you requested, the correspondence from Professor Irving Fisher, who is much more intelligent but certainly misled on this point" (*MSE*, Eccles to M. H. McIntyre, letter, June 2, 1938). "If he felt he was right," recalls Currie, "he was not in the slightest impressed by 'authority' to the contrary" (Currie to author, letter, August 24, 1983). This characteristic was evident in his belief, contrary to the opinions and advice of

his economists, that the war would be followed by inflation, not deflation.

Quite aside from any influences from economists to which Eccles might have been subjected during the early 1930's or later, the fact remains that he came to a view of the workings of the macroeconomy which was practically unthinkable for one from his background. Eccles attributed the advanced nature of his thinking to the fact that he was a country banker and had not attended college (Stein, p. 485, fn. 47). Currie's explanation is to suggest that "he was what in biology is called a mutation!" (Currie to author, letter, August 24, 1983). Perhaps Eccles' Mormon background preconditioned him to see a role for government in economic planning and a stabilization. Mormon economic policy during the nineteenth century was characterized by strong central direction and control, and Mormon economic institutions blended principles of self-reliance and cooperation.

Whatever the explanation for the evolution of Eccles' ideas on macroeconomic policy, his was a remarkable intellectual accomplishment, and one which has had lasting impact. Musgrave, Eccles' personal assistant from 1944 to 1948, considers Eccles as "a great figure in a crucial period of our history... an extraordinary figure, with a great deal of insight and courage. ...It is easy," writes Musgrave, "to belittle theoretical qualities in a man of action such as Eccles; yet which academic, other than Keynes, was as important in implementing a modern view of macro policy into actual policy measures?" (Musgrave to author, letter, September 22, 1983). Friedman does not exempt even Keynes from the assessment when he states,

"I believe [Eccles] played a far greater role in the development of what came later to be called Keynesian policies than did Keynes or any of his disciples" (Friedman to author, letter, July 22, 1983).

Eccles was a paradox: a developer of economic theories who expressed an intense dislike of "theory" and categorically denied being an economist; a "conservative Republican businessman-banker" who served a Democratic president and delighted in pointing out logical flaws and backwardness in the thinking of bankers and businessmen; and a defender of the free-enterprise market economy who believed that only significant government intervention in the economy could save the system.

## REFERENCES

- Arrington, Leonard J., *David Eccles, Pioneer Western Industrialist*, Logan: Utah State University Press, 1975.
- Hyman, Sidney, *Beckoning Frontiers, Public and Personal Recollections*, New York: Alfred Knopf, 1951.
- \_\_\_\_\_, *Marriner S. Eccles, Private Entrepreneur and Public Servant*, Stanford: Graduate School of Business, Stanford University, 1976.
- May, Dean L., *From New Deal to New Economics, The Liberal Response to the Recession*, New York: Garland Publishing, 1981.
- Stein, Herbert, *The Fiscal Revolution in America*, Chicago: University of Chicago Press, 1969.
- Marriner S. Eccles Collection (MSE)*, Special Collections Library, University of Utah, Salt Lake City.

# Rudolf Hilferding: The Dominion of Capitalism and the Dominion of Gold

By WILLIAM A. DARITY, JR. AND BOBBIE L. HORN\*

In November 1918, the USPD-SDP provisional revolutionary government established a commission to study the socialization of German industry. Its members included Karl Kautsky, Emil Lederer, and, of all people, Joseph Schumpeter. Later when asked how he could have been connected with such a commission, Schumpeter reportedly replied that if somebody wants to commit suicide, it is a good thing if a doctor is present. Weimar Germany had its "doctor," a trained physician, also a member of the socialization commission, in Rudolf Hilferding. As the economic expert of the SDP and twice Finance Minister in coalition governments, Hilferding sought to prescribe treatments as German socialism's humanitarian midwife.

Hilferding's name probably is most familiar to economists for his early skirmish with the dreaded Böhm-Bawerk over the equally dreaded transformation problem and for the qualified recognition of his treatise *Finance Capital* in Lenin's *Imperialism*. A number of factors should stimulate a renewed interest in this early theoretician of corporate capitalism. First, the recent publication of an English translation of *Finance Capital* (1981) means fewer will be limited to accepting authoritative summaries without some perusal of its contents. Second, there is a continuing and important interest in Marx's theory of money. Third, finance and financial crises persist in intriguing economists and still await satisfactory treatment. Combined with the receding tides of Keynesianism and Monetarism, and the resurgence of "Classicals" of all colors, it may prove interesting to reconsider Hilferding's efforts.

In this paper, we address three issues. First, we look at Hilferding's career and his contributions to Austro-Marxist doctrine. Second, we consider his two experiences as Finance Minister: 1) facing the crisis of the infamous German hyperinflation; 2) pursuing deflationary policies on the eve of the Great Depression. Last, we offer some very brief comments on the relationship between his political activities and his intellectual work as a socialist theoretician.

## I. Austro-Marxism

Born into what Paul Sweezy described as a "well-do-do Jewish mercantile family" (1949, p. xv) in Austria in 1877, Hilferding trained in medicine at the University of Vienna, receiving his doctorate in 1901. While at the university he became actively involved in the socialist movement and was drawn to the study of political economy. His contact with the central figures of the "Austrian school" laid the foundation for his Marxist critique of subjectivist approaches in economics.

Hilferding soon became a leading figure in the Austro-Marxist school. In 1904 with Max Adler, he inaugurated the *Marx-Studien*, the theoretical organ of the Austro-Marxists. In 1906 he was lecturing at the Workers University at Berlin, along with Rosa Luxemburg. He subsequently edited *Vorwärts*, a major socialist newspaper and in 1910 published his major work *Finance Capital: A Study of the Latest Phase of Capitalist Development*. After the war he edited *Freiheit*, the journal of the Independent German Social Democratic Party (USPD) and upon the fragmentation of the USPD moved with other right wing independents to the German Social Democratic Party (SDP). He obtained German citizenship in 1920 and served as Weimar's Minister of Finance briefly in 1923 and again, for a somewhat more extended

\*University of North Carolina, Chapel Hill, NC 27514, and University of Tulsa, Tulsa, OK 74104. We thank Steve Steib, Cadwell Ray, and Charles Kindleberger for comments.

period, in 1928–29. While not holding a cabinet post, Hilferding served as a member of the Reichstag and edited the SDP's theoretical monthly, *Die Gesellschaft*. With the victory of the National Socialists—a possibility denied by Hilferding only a few days before Hitler's appointment to the Chancellorship—he was compelled to flee into exile and forced to redirect his attention from the Communist threat to the harder reality of the National Socialists. Eventually turned over to the Gestapo by the French authorities in 1941, he reportedly hung himself in his jail cell.

Hilferding is considered the leading economic theoretician of Austro-Marxism, a movement that is of special importance to economists because of a common historical and intellectual environment shared with the Austrian school of economics. Both Austro-Marxism and Austrian economics reflected similar influences and demonstrated mutual awareness of their respective approaches. Hilferding displayed sensitivity to the Austrian tradition (along with the German monetary tradition) in the pages of *Finance Capital* and felt obliged to defend Marx from critiques launched from such perspectives. Both Hilferding's Austro-Marxist economics and that of the Austrian school placed special emphasis on distortions in the structure of prices as fundamental to the propagation of capitalist crises. Hilferding's interesting attitude toward the price system is apparent in his last publication (1963). However, for the Austrians the source of these distortions was autonomous to the market process, coming particularly from the state. In contrast, for Hilferding these distortions had an endogenous origin in the normal workings of a capitalist economy, even under thoroughgoing *laissez-faire*, though the state could be one source.

Hilferding's Austro-Marxist economics produced two related theories of capitalist crises. The primary theory was his conversion of Marx's "law of the tendency of the rate of profit to fall" into simply a statement about the capitalist business cycle. Hilferding, in rejecting what he termed "the dogma of the falling interest rate" (1981, p. 104), shifted to a treatment where the rate of profit

is something that swings periodically rather than secularly downward. Thus, he moved away from Marx's conceptual interpretation of the rate of profit in pure form. This move from Marx's original approach may have been due to Hilferding's inclination to view Marx's work as the culmination of classical political economy rather than the most substantive critique advanced to that date (Henryk Grossman, 1977, p. 48). For better or worse, Hilferding's Austro-Marxist economics was a variant of Marxism, without immunity to influences from Austrian economics or other "modern" movements—an idiosyncratic species of Marxism indeed!

While Hilferding viewed capitalism as historically doomed, he did not (as did Marx) attribute its demise directly to the tension between the progressive reduction of socially necessary labor time and the fact that labor power constituted the sole source of profit. Hilferding foresaw a transformation of the economy with growing centralization and concentration of capital as the normal outcome of competition under capitalism.

Hilferding's second major explanation for capitalist crises, in addition to cyclical swings in the rate of profit, was the classic Volume II problem of disproportionality in production. The failure of the capitalist pricing mechanism to produce the appropriate signals would lead to imbalances in the production of goods across the various sectors—especially between wage goods and capital goods. In fact Hilferding even suggested that the cyclical drop in the rate of profit would go hand in hand with "the disruption of these proportional relations." (For a complete discussion, see pp. 239–66.)

In exploring the crystallization of crisis, Hilferding identified three primary factors in the recurrent fall in the profit rate: 1) the extension of turnover time, 2) the rise in the wage rate associated with the growth in demand during the upturn of the cycle, and 3) the rise in the rate of interest above normal levels adversely affecting entrepreneurial profit. However, somewhat strangely, the crisis appeared to lack inevitability: "A monetary crisis is not an absolutely necessary feature of the crisis, and may not always occur" (p. 274). Banks could avert the crisis

if they would continue to make credit available to producers. But the private banks could, not or would not, make funds available for two major reasons, according to Hilferding:

In the first place, speculation in both commodities and securities is in full spate and makes increasing demands on the supply of credit. Second,..., the circulation credit which producers extend to each other becomes inadequate to meet the increased demands, and here too the banks must help out.

[p. 260]

The banks would do their utmost to keep their retained profits "in liquid form, as money"; therefore the conversion into "productive capital," that brings high employment and continued prosperity, does not take place. A Hilferding crisis of economic depression, although finding its origins outside of the specific characteristics of the financial apparatus, is often confirmed by the withdrawal of credit. It is in keeping with Hilferding's position, as the economic theoretician of the Austro-Marxist school, to construct such a vision of capitalism—unstable, volatile, prone to crises—the antithesis of the vision of Austrian economics, yet sharing many of the same views about the mechanisms of the capitalist economic system.

## II. The Socialist Finance Minister

Hilferding's brief first tenure as Finance Minister began in August 1923, at the height of the hyperinflation, when Gustav Stresemann was called upon to form a coalition government. Hilferding clearly inherited a situation which called for immediate and radical action. Hilferding had the opportunity to adopt Lenin's dictum (as reported by Keynes, 1963, p. 77) and attempt to "destroy the Capitalist System" by debauching the currency. But Hilferding did not want to destroy capitalism; he was not looking for an economic collapse, but rather "a collapse which will be political and social..." (1981, p. 366), amounting to a social and political transformation of the capitalist economy. For Hilferding, the pressing problem was that of

domestic inflation. Renegotiation on reparations, and almost all else, required monetary stabilization as a precondition.

It is difficult, if not impossible, to gauge the nature and extent of the gulf between the theory of *Finance Capital* and Hilferding's policies as Finance Minister. One would assume that Part I of *Finance Capital*, "Money and Credit," would be relevant to his political practice, particularly in an era of hyperinflation. For some of his critics, it was relevant; that was one of Hilferding's problems.

Part I is probably the most controversial portion of the book. It was about this section that Lenin expressed the reservation that *Finance Capital* "gives a valuable theoretical analysis" despite an unspecified "mistake the author commits on the theory of money" (p. 11). Even Hilferding's theoretical ally Kautsky (see Harold James, 1981, p. 852) voiced disagreement, and Schumpeter dismissed it as "rather old-fashioned monetary theory" (1954, p. 881). So too with modern commentators: The editor of the new English translation describes it as "[p]erhaps the least successful part of the book" (p. 5); reviewer Anthony Brewer (1983, p. 102) suggests skipping the section; and David Harvey (1982, pp. 290–92), following de Brunhoff, situates many of Hilferding's errors in his misreading of Marx on money.

One of the primary tasks in Hilferding's (1981) treatment of money is to offer an explanation of the determination of the value of a pure, nonredeemable, state paper money; an explanation of the "modern monetary experience" of Holland, Austria, and India.

Hilferding's monetary doctrines do not rest comfortably in any of the standard monetary camps. He accepts that "ever since Tooke's demonstration, the quantity theory of money has been rightly regarded as untenable" (p. 47); however, "there is a reluctance to give due recognition to the influence of quantity on the value of money even where it really is the determining factor, as in the case of paper money and depreciated currency" (p. 50).

From Hilferding's perspective, the theoreticians of the Cuno government, the Knapp-Helfferich school, simply went too far in their rejection of the quantity theory and

were stymied by their value theory from devising a theoretically satisfactory explanation. For Hilferding, "The quantity theory, then, holds good for a currency with suspended coinage. After all, the theory was formulated as a generalization of the experience with unsettled currencies at the end of the eighteenth century in America, France and England" (p. 55).

He gives a number of reasons for the impossibility of a pure paper currency system. "A pure paper currency is, therefore, impossible as a permanent institution, because it would subject circulation to constant disturbances" (p. 57). However he lays out a system in the "abstract" in which the quantity of "legal tender state paper money" could not be increased, and "[t]he impossibility of increasing the supply of paper money would protect it against depreciation" (p. 57), while banks could provide an endogenous credit money component to prevent appreciation. "Under such circumstances, paper money would behave as gold does today..." However, "reality" throws up three obstacles to such a scheme:

In the first place, this paper money would be valid only within the boundaries of a single state. For settlement of international balances, metallic money with an intrinsic value would be required; and if this requirement is to be satisfied, the value of the money in domestic circulation must be kept on a par with the medium of international payments to avoid the disruption of commercial relations. [pp. 57-58]

Second, "there is no possible guarantee that the state will not increase the issue of paper money." He ends with a third reason: "money with an intrinsic value—such as gold—is always needed as a means of storing wealth in a form in which it is always available for use." For Hilferding, explanation of Germany's inflation was straightforward.

Hilferding's analysis would require linking of the mark to gold, primarily to constrain the state's financial activities. But also, for him capitalism maintains a certain infatuation:

Credit collapses, and thus suddenly deserted capitalism returns in despair to its first love, to gold.... Capitalism thought that it had long since liberated itself from the domination of gold, but now it experiences a bitter disillusionment, and shaken by panic reorganizes its continuing dependence. But such crises are cathartic.... Nevertheless, the more capitalism succeeded in establishing its own domination, the less did it allow itself to be bound by this golden chain. The loved one, once so demanding, learns to be more modest and is eventually satisfied with the role of someone in reserve.... Her demands may become excessive, and she may occasionally refuse her favors altogether, but these moods do not last long and things soon return to normal. Gold has lost, once and for all, its absolute domination.... [p. 274]

He saw that even the authority of the state was limited ultimately under the laws (and loves) capitalism. In October of 1923, Keynes was writing:

A regulated non-metallic standard has slipped in unnoticed. *It exists*. Whilst the economists dozed, the academic dream of a hundred years, doffing its cap and gown, clad in paper rags, has crept into the real world by means of the bad fairies—always so much more potent than the good—the wicked Ministers of Finance. [pp. 208-09]

The Minister Hilferding, perhaps not sufficiently wicked, accepted limits on the powers of even the bad fairies.

To most interests, any return to gold appeared strongly deflationary in the context of the German situation. But in Hilferding's theory it was not necessarily the case. For with a stable money, the banking system could create a sufficient quantity of domestic credit money to maintain the "social minimum of circulation." But, moreover, in spite of any deflationary implications, there appeared no alternative.

Opponents to the return to gold possessed as one alternative, Helfferich's "ryemark/

rentenmark" scheme, which appeared to avoid the deflationary implications of returning to gold while instilling public confidence in a new currency; partly a scheme, partly a bluff (see Erich Eyck, 1963).

Hilferding's role in the ultimate action is unclear. He served as Finance Minister from mid-August to the end of September, and the currency reform was introduced in mid-October after Hilferding had been forced to resign. The ultimate plan adopted was a hybrid of the two proposals with a Rentenbank being established, issuing notes backed by mortgages with the value not tied to rye, but instead to gold. The other crucial aspect was a restriction on the Reich in terms of discounting its bills with the Reichsbank.

Hilferding returned to the post of Finance Minister in June of 1928 in the coalition government of Hermann Muller. This year marked the last of the "golden years" with unemployment of 7 percent, just before the massive collapse driving unemployment beyond 30 percent by 1932. Hilferding remained in office until December 1929 when he resigned in a controversy surrounding the negotiations of a state loan with an American banking concern. The primary concern of Hilferding in this period was the control of the budget and arrangements to "fund," as opposed to monetize, that portion that could not be covered by taxes. The deficits from the budgets of 1926 and 1927 had created funding problems and the unemployment insurance credits were rapidly mounting. Hilferding proposed both budget reductions and tax increases, policies that met with wide opposition, all the while resisting "inflationist" schemes.

So, though not necessarily a "hard-money" man, Hilferding was a "sound-money" one. His position seems remarkably close to that of an individual who politically could not be more different, an American contemporary, J. Lawrence Laughlin, who also thought of himself as a monetary doctor of sorts. Both rejected the quantity theory, except for the case of inconvertible state paper money; and, moreover, both rejected the viability of such an institution as a permanent arrangement. Both accepted some form of the so-called "real bills" doctrine and possessed some form

of "objective" value theory. Theoretical consistency and policy construction often produces peculiar bedfellows.

### III. Theory and Practice

Given Hilferding's notion of the increasing concentration and centralization of finance capital, an idea that ultimately matured into his concept of "organized capitalism," he seems to have thought that it would be better to simply let the system evolve, according to its own inherent logic. Let the system develop organizational forms and mechanisms to cope with the surface manifestations of the underlying contradictions, and then simply take it over, by democratic means, with its essentially "socialistic" organizational structure already in place.

Analysis of Hilferding's theoretical stance is both complemented and complicated by the fact that his political activity seems to have offered the opportunity to test his theory. It is a widely held view, particularly on the left, that, to quote Sweezy: "his record, like that of the Social Democratic Party itself, was one of unbroken failure. As Finance Minister he was equally ineffective in dealing with inflation in 1923 and with impending depression in 1929" (pp. xvii-xviii).

Gerd Hardach et al. echo this view and write that Hilferding "made drastic mistakes on virtually every relevant economic question of the time: on structural unemployment in the 1920s, on the outbreak of the world crisis and finally on stabilization policy" (1978, p. 56). These authors attribute these "drastic mistakes" to Hilferding's "...complete reformism. Marxist theory was...only a rhetorical reference point—for concrete analysis social democracy relied on bourgeois economics" (p. 56).

Hilferding seemed to view Marxian political economy as having an almost positivistic scientific autonomy.

...[S]o far as Marxism is concerned the sole aim of any inquiry—even into matters of policy—is the discovery of causal relationships.... According to the Marxist conception, the explanation of how such class decisions are



determined is the task of a scientific, that is to say a causal, analysis of policy. The practice of Marxism, as well as its theory, is free from value judgements.

[p. 23]

This view of Marx and this separation of Marxist political economy from some inherent working class perspective can be sensed from Schumpeter's evaluation of Hilferding's as Finance Minister:

The minister Hilferding, much too good an economist not to see what was wrong and much too good a Marxist not to realize that there are situations in which anticapitalist policy is in the end anti-socialist, actually went so far as to attempt a very 'capitalistic' fiscal reform. [1939, p. 715]

#### REFERENCES

- Brewer, Anthony, "Review of *Finance Capital*," *Economica*, February 1983, 50, 102-03.
- Eyck, Erich, *A History of the Weimar Republic*, Cambridge: Harvard University Press, 1963.
- Grossman, Henryk, "Archive: Marx, Classical Political Economy and the Problem of Dynamics," *Capital and Class*, 1977, 2, 32-55.
- Hardach, Gerd, Karras, Dieter and Fine, Ben, *A Short History of Socialist Economic Thought*, New York: St. Martin's Press, 1978.
- Harvey, David, *The Limits to Capital*, Oxford: Oxford University Press, 1982.
- Hilferding, Rudolf, *Finance Capital: A Study of the Latest Phase of Capitalist Development*, in Tom Bottomore, ed., London: Routledge and Kegan Paul, 1981.
- , "State Capitalism or Totalitarian State Economy," in C. Wright Mills, ed., *The Marxists*, New York: Dell Publishing, 1963.
- James, Harold, "Rudolf Hilferding and the Application of the Political Economy of the Second International," *Historical Journal*, 1981, 24, 847-69.
- Keynes, J. M., *Essays in Persuasion*, New York: W. W. Norton, 1963.
- Lenin, V. I., *Imperialism, the Highest Stage of Capitalism*, Peking: Foreign Languages Press, 1975.
- Schumpeter, Joseph A., *Business Cycles*, Vol. II, New York: McGraw-Hill, 1939.
- , *History of Economic Analysis*, New York: Oxford University Press, 1954.
- Sweezy, Paul, *Karl Marx and the Close of His System by Eugen von Böhm-Bawerk & Böhm-Bawerk's Criticism of Marx by Rudolf Hilferding*, New York: Augustus M. Kelley, 1949.

# Korekiyo Takahashi and Japan's Recovery from the Great Depression

By DICK K. NANTO AND SHINJI TAKAGI\*

The economic policies of Japan's Finance Minister, Korekiyo Takahashi, during the Great Depression are more than a curious footnote in economic history. By 1932, Takahashi's interventionist policies had reversed Japan's economic decline. They also preceded similar measures adopted in other countries. Takahashi has even been characterized as Japan's Keynes, because of his use of deficit-financed, fiscal stimuli. He also has been blamed for his role in initiating in Japan the process of expanding military expenditures to stimulate the economy.

This paper takes a second look at the Japanese economy during the early years of the Great Depression, with a particular reference to the policies of Takahashi. Two major questions addressed relate to the Keynesian origin of his ideas and the extent to which his policies contributed to the subsequent economic recovery.

## I. Economic Background

In real terms, the Great Depression scarcely occurred in Japan. From 1926 through the 1930's, real gross national product (*GNP*) in each succeeding year never declined from its previous level. The economy did experience some stagnation in the 1929–31 period; real *GNP* rose by only 0.5 percent in 1929, 1.1 percent in 1930, and 0.4 percent in 1931. For the remainder of the 1930's, however, real *GNP* grew at an average annual rate of 5.8 percent.

Real gross domestic product (*GDP*) by sector, however, did decline from peak to

trough during the 1928–33 period; it fell by 7 percent overall, 12 percent for agriculture, 7 percent for construction, 6 percent for transportation and communications, and 25 percent for commerce and services. For manufacturing and mining, however, real gross domestic production rose in each year of the Great Depression except for 1932, when it dropped by 1 percent.

What did occur during this time were severe upheavals in Japan's monetary and fiscal systems in concert with worsening economic conditions abroad. This caused overall stagnation, declines in production in some sectors, a severe deflation, and rapid losses in Japan's gold reserves. This was compounded by major crop failures and sharp reversals in monetary and fiscal policies.

Declines in nominal *GNP* levels defined the dimensions of the depression for most producers. The *GNP* dropped by 1 percent in 1929, 10 percent in 1930, and 9 percent in 1931, before beginning to recover. The declines were caused, however, mostly by a drop in prices and not in production.

In Japan, the episode of the Great Depression was preceded by a decade of general economic stagnation following the wartime boom of 1914–18 and was precipitated by a series of political and economic events that began with the financial panic of 1927.

In 1927, following the closing of the Bank of Taiwan, bank runs swept the nation. By the time a three-week moratorium was declared, as many as 27 banks had suspended operations. The Bank of Japan came to the rescue by opening its discount window and nearly tripling its loans in one month. As the crisis ended, banks were awash in liquidity but reluctant to make risky loans.

As the economy continued to perform sluggishly, pressures began to build to return to the gold standard. Bankers wanted to be able to invest their excess liquidity overseas,

\*Congressional Research Service, Library of Congress, Washington, D.C. 20540, and International Monetary Fund, Washington, D.C. 20431, respectively. Views expressed do not necessarily reflect those of the International Monetary Fund.

while the Ministry of Finance was facing the need to refinance 230 million yen in government bonds held abroad (Kozo Yamamura, 1972, p. 195).

Under these circumstances, the Kenseikai party formed a new cabinet in July 1929 with its Finance Minister, fiscally conservative Junnosuke Inoue, who had long advocated returning to the gold standard. On January 11, 1930, just three months after the Great Crash on Wall Street, Japan went back to the gold standard at the higher pre-World War I parity for the yen.

In retrospect, there could not have been a worse time to return to the gold standard. Japan was hit by a modern-day gold rush never thought possible by the Kenseikai. In 1930, 309 million yen in gold left Japan, and the money supply fell by 13 percent. Prices dropped by 10 percent, causing the real money supply to fall by 2.5 percent, and production in certain sectors to decline. Compounding these problems was a fall in foreign demand for Japanese exports of silk and cotton.

Despite these worsening economic developments at home and abroad, the government and financial leaders still favored the gold standard as late as mid-1931. Attitudes began to change, however, with the collapse of the European financial market in July, a deteriorating military situation precipitated by the Manchurian Incident in September, and the abandonment of the gold standard by England in September.

Under these circumstances and early into the Great Depression, the Seiyukai party regained power and called Takahashi, then 77, back from retirement to serve as Finance Minister for the fifth time.

## II. Korekiyo Takahashi—The Man

Korekiyo Takahashi was born in 1854, the year Commodore Perry sailed his black ships into Tokyo Bay, and died under the weapons of military assassins in 1936; the year that Keynes' *General Theory* was published. Seven times appointed Finance Minister, Takahashi was one of Japan's premier Ministers of Finance and molders of economic policy (Kiyoshi Oshima, 1969).

Adopted into a low-ranking samurai family, Takahashi learned English and in 1867 was sent to the United States for further study. There, contrary to an initial agreement, he ended up as an indentured servant. Rescued from this servitude through the intervention of an American friend, he returned to Japan in 1868 after hearing of the Meiji Restoration.

Takahashi became the first Director of the Patent Office in 1884 and later went to Europe and the United States to study patent law. After a failed silver mining venture in Peru, he joined the Bank of Japan in 1892 where his assignments included selling war bonds in the United States and Europe during 1904–07. In 1911, at the age of 57, he became Governor of the Bank of Japan.

After reaching the top at the Bank of Japan, new horizons beckoned Takahashi. In 1913, he joined the Seiyukai and was soon appointed Finance Minister. Over the next fifteen years, he served in that capacity under four different cabinets, including one of his own. When he was called back to handle the 1927 financial panic, he was already over 70-years-old and in semiretirement. He immediately declared a 21-day moratorium on payments of debts, rode out the panic, and resigned after only 42 days in office.

The 1927 panic, however, was only a harbinger of things to come. In December 1931, as Japan was slipping into the Great Depression, he was appointed Finance Minister for the fifth time and was destined to serve under two more cabinets until his death on February 26, 1936.

## III. Takahashi's Economic Policies

Takahashi lost little time in reversing the contractionary economic policies of the previous cabinet when he resumed office at the end of 1931. The result has been described as "one of the most successful combinations of fiscal, monetary, and foreign exchange rate policies, in an adverse international environment, that the world has ever seen" (Hugh Patrick, 1971, p. 256). The recession that had begun in 1930 was reversed by late 1932, and Japan escaped most of the ravages of the Great Depression.

The essence of Takahashi's economic policies included: 1) the abandonment of the gold standard (December 1931) and export promotion; 2) low interest rates; and 3) deficit spending financed by the sale of bonds through the Bank of Japan (beginning with the fiscal 1932 budget) (Shiro Mori, 1975).

#### A. *The Gold Standard*

The decision to go off gold likely would have been made regardless of who was Finance Minister. Japan probably would not have continued much longer on the standard after England had already abandoned it, after Japan's reserves of gold had dropped in two years from 1,343 million to 557 million yen, and after much of the business community had turned against it.

Still the decision fit into Takahashi's view of the world. He was highly nationalistic and somewhat of a mercantilist. He had opposed the gold standard in the early 1920's and felt Japan's hard-earned gold reserves should be retained—partly to finance colonial activities on the Asian continent and partly as a reserve against future contingencies. He also opposed the attempt to raise the exchange value of the yen.

After the yen's ties to gold were severed, the exchange rate with the dollar depreciated by 50 percent in 1932 and another 50 percent in 1933 before recovering somewhat in 1934. For the four years under Takahashi, the exchange value averaged 4.0 yen as compared with 2.2 yen per dollar for the four years prior to his becoming Finance Minister.

The devaluation of the yen was allowed in order to stimulate exports. As the economy was stimulated, exchange controls were imposed to keep imports from rising too fast. Such exchange controls preceded those adopted by the United States, England, and Germany (Takafusa Nakamura, 1983, p. 233).

The result was a dramatic rise in Japanese exports of goods and services both in nominal and real terms. Nominal exports, which had dropped by 25 percent in 1930 and 16 percent in 1931, rose by 22 percent in 1932 and 25 percent in 1933. In real terms, ex-

ports, which had risen by only 7 percent over the previous two years, jumped by 19 percent in 1932 and 6 percent in 1933. The current account surplus accounted for 38 percent of the rise in real *GNP* over the 1932–33 period.

#### B. *Low Interest Rates*

Takahashi had been a long-time advocate of low interest rates in order to promote business investment and expansion. He had been criticized after World War I for helping inefficient industries through such policies and postponing their needed adjustment. His experience as Governor of the Bank of Japan and close ties to the business community reinforced his adherence to this policy. Low interest rates also were needed to sell government bonds and finance the debt.

Takahashi, furthermore, linked interest rates to income distribution and consumption demand. He thought high rates reduced the share of income going to labor, which reduced the purchasing power of the people, lowered demand, and caused recessions. A low interest rate policy, he argued, was the most effective policy tool to counteract a recession (1936c, p. 235).

Under Takahashi, the Bank of Japan lowered its discount rate from 5–6 percent to about 3 percent (by 1935). Other interest rates followed this downward trend. In 1932, the real call-money rate (a market rate) likewise dropped from about 15 percent to 6 percent and remained at less than 1 percent for the remainder of the Takahashi years.

Partly as a result of this policy, real gross domestic fixed capital formation recovered to its 1929 peak by 1933 and rose steadily thereafter.

#### C. *Deficit Spending*

Until 1932, when Takahashi became Finance Minister, Japanese government borrowing for deficit spending had been done for specific purposes, such as public works, reconstruction, or war. Under Takahashi, however, funds from new debt issued could be used for general government expenditures. Government debt, moreover, was allowed to be financed by selling bonds to the

Bank of Japan, which then resold them to the public. This constituted the beginning of "open market operations" in Japan, a term Takahashi (1936a, p. 579) used in its English form.

Government fiscal policy, which was turning expansionary in Inoue's last year, then switched completely from attempts to cut the budget and reduce debt to deficit-financed spending. Net central government expenditures, which had fallen by 19 percent since their peak in 1928, rose by 32 percent in 1932 and another 16 percent in 1933. Over these two years, increased government consumption and capital formation accounted for 29 percent of the total increase in real GDP.

Of course, some of the rise in government spending was a natural consequence of pressures by the military and emergency expenditures for local relief for farmers because of crop failures. Of the large increases in government expenditures in 1932 and 1933, approximately half were accounted for by increased military spending. By 1935, 78 percent of the total rise in central government expenditures had been in the military accounts. Relief for farm villages also was important, but the bulk of these expenditures came from local jurisdictions, not from the central government.

#### IV. The Keynesian Origin of Takahashi's Policies

The expansionary economic policies advocated by Takahashi were similar to those he had pursued during previous recessions. During his later years, however, his justification for them and his understanding of their impact on the economy seemed to have acquired an increasingly Keynesian nature.

In terms of foreign exchange policy, Takahashi thought the yen's value should be set by the market (1936a, p. 423). Like Keynes, he opposed going back to the gold standard at the prewar parity, and he considered internal price stability to have precedence over external balance (Keynes, 1923, pp. 140-76).

Takahashi's low interest rate policy was Keynesian in the sense that it was based on

the link between the monetary and the real sectors. His perception of this link was not as complete as that in *The General Theory*, but he saw changes in monetary conditions affecting real output through changes in wages and interest rates. Takahashi was skeptical of the quantity theory of money as a basis for policy (1936b, pp. 217-18).

Takahashi justified increased government expenditures, especially for the military, in terms of the expenditure multiplier and the business activity they generated. In October 1929, he demonstrated his knowledge of the Keynesian multiplier in an explanation of what is now considered to be the paradox of thrift. He noted that if an individual were to save new income instead of spending it, the multiplier would be 1. If the money were spent, however, he thought the multiplier would be as high as 20 to 30 (1936c, p. 247).

The fact that Takahashi justified government spending to stimulate employment and economic activity in terms of the multiplier, however, does not mean, as many have argued, that his ideas predated Keynes, or, as one author has postulated, that he could not have given the speech in 1929 before Keynes published his *General Theory* (Shogo Sasahara, 1981, p. 225). The distinctive contribution of *The General Theory* was not the advocacy of a policy of public works expenditure. Such a policy had already been argued by Keynes in 1929 in an influential pamphlet coauthored with Hubert Henderson entitled, *Can Lloyd George Do It?* (Don Patinkin and J. C. Leith, 1977, p. 9). This pamphlet not only explained the difference between primary and secondary effects of government spending, but laid out the idea of the expenditure multiplier.

How Takahashi acquired the concept of the expenditure multiplier is not known, but it is quite likely that he had read the Keynesian description of it by 1929. He was an avid reader of foreign books and periodicals, including the *London Times* (Ministry of Finance, 1977, p. 128). He kept abreast of theoretical developments in economics abroad and used them in his public statements. In a 1933 speech, for instance, he quoted Irving Fisher and used the results of a University of Chicago economic seminar

that he had translated in arguing for a countercyclical fiscal policy (1936a, p. 559). He also advocated several ideas which had appeared in Keynes' earlier writings on money.

Takahashi, however, understood that financing government expenditures by issuing debt was limited by the ability of the public to absorb it. Beyond that it becomes inflationary. In Takahashi's words, "the limit comes when the effect of the additional funds raised by issuing debt has no value in fostering private industry and, therefore, no stimulus on sound development" (1936a, pp. 54-55.) He saw this limit being reached in 1935 when he made the fatal decision to attempt to rein in military spending.

### V. Factors Underlying the Economic Recovery

Regardless of the origin of the ideas behind the policies advocated by Takahashi, it is of independent interest to assess their quantitative impact on the subsequent economic recovery that started towards the end of 1932.

Since the beginning of the recovery coincided with the revision of the government budget, and the subsequent expansion paralleled the continuation of the budgetary deficit in succeeding years, the contribution of Takahashi's fiscal policies to the economic recovery has not been questioned much in Japan. Recently, however, some have noted the strong recovery in private investment during the early phase of the recovery and suggested that the economic expansion was actually led by the private sector (K. Shima, 1983, pp. 108-13).

In order to test the effect of various economic variables on real *GDP*, we use Granger causality tests (C. W. J. Granger, 1969; R. J. Gordon and J. A. Wilcox, 1981). These were done by using an *F*-test to determine whether changes in real *GDP* can be explained by variables other than the lagged value of real *GDP*, itself. The tests were performed by using annual data for the period 1920-40, with the current and two lagged values of several independent variables, of which five were significant. Since only annual data could be employed, current values were also included to capture the feedback process dur-

TABLE 1—GRANGER CAUSALITY TESTS ON JAPAN'S REAL *GDP*, 1920-40<sup>a</sup>

Independent Variable	<i>t</i> -Ratio Current Value	<i>F</i> -Statistics	
		Two Lagged Values	Current and Lagged Values
Consumer Price Index	2.144 <sup>c</sup> (1.608)	2.656 (1.447)	1.771 (0.966)
Export Price Index	3.304 <sup>b</sup> (2.478) <sup>b</sup>	4.009 <sup>b</sup> (1.447)	3.660 <sup>b</sup> (2.179)
Yen/Dollar	1.958 <sup>c</sup>	2.784 <sup>c</sup>	2.119
Exchange Rate	(1.942) <sup>c</sup>	(0.499)	(1.631)
Real Central Government Expenditures	1.641 (1.621)	1.455 (0.774)	4.930 <sup>b</sup> (4.016) <sup>b</sup>
Real Domestic Debt	1.921 <sup>c</sup> (0.790)	0.322 (0.130)	3.316 <sup>c</sup> (3.601) <sup>b</sup>
Real Private Investment	1.683 (1.245)	1.933 (0.432)	1.633 (0.733)

Sources: Underlying data from Ohkawa et al. (various years); Bank of Japan, *Economic Statistics of Japan*, annual issues; *Oriental Economist Yearbook*, annual issues.

<sup>a</sup> The results for money supply, interest rates, and real exports were insignificant and not reported. Although insignificant, the results for real private investment are reported in view of the current debate on its role in the recovery. Figures based on log and level (in parentheses) are shown.

<sup>b</sup> Significant at 5 percent.

<sup>c</sup> Significant at 10 percent.

ing the same time period. Summary statistics from these tests are presented in Table 1.

Note that real private investment had little causal (lagged) as well as contemporaneous (current) impact on real *GDP*. Although, not surprisingly, real central government expenditures and real domestic debt did have a significant contemporaneous impact, they had little causal effect on real *GDP*.

What did have significant causal as well as contemporaneous effects were external price developments, captured by exchange rate and export price movements. It was also found that, in terms of the impact elasticity of the current value, the consumer price index (0.569) had the strongest impact on real *GDP* of all the independent variables.

The above results, however, are subject to the usual interpretational limitations associated with Granger causality tests. Moreover, they are more limited by the use of a relatively long, annual time-series. Use of

monthly or even quarterly data might provide a better basis for interpreting the causal relationships observed among different economic variables, if such time-series were available.

Given these limitations, however, the importance of price factors—primarily brought about by Takahashi's exchange rate policies—in the subsequent economic recovery is clearly indicated. On the other hand, the contribution of fiscal policies, as important as they may be from the point of view of the history of economic thought in Japan, may well have been given too much credit.

# REFERENCES

- Gordon, R. J. and Wilcox, J. A., "Monetarist Interpretations of the Great Depression," in K. Brunner, ed., *The Great Depression Revisited*, Boston: Martinus Nijhoff, 1981.
- Goto, Shinichi, *Takahashi Korekiyo, Nihon no "Keynes,"* Tokyo: Nihon Keizai Shimbun, 1977.
- Granger, C. W. J., "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods," *Econometrica*, July 1969, 37, 424–38.
- Keynes, John M., *A Tract on Monetary Reform*, London: MacMillan, 1923.
- Mori, Shiro, "Fiscal Policy of Korekiyo Takahashi," *Shokei Ronso*, 1975, 10, 1–42.
- Nakamura, Takafusa, *Economic Growth in Pre-war Japan*, New Haven: Yale University Press, 1983.
- Ohkawa et al., Kazushi, *Estimates of Long-Term Economic Statistics of Japan since 1868*, Vols. 1, 7, 8, Tokyo: Toyo Keizai Shinposha, 1966, 1967, 1974.
- Oshima, Kiyoshi, *Takahashi Korekiyo*, Tokyo: Chuo Koronsha, 1969.
- Patinkin, Don and Leith, J. C., *Keynes, Cambridge and The General Theory* London: Macmillan, 1977.
- Patrick, Hugh, "The Economic Muddle of the 1920's," in J. W. Morely, ed., *Dilemmas of Growth in Prewar Japan*, Princeton: Princeton University Press, 1971.
- Sasahara, Shogo, "Theoretical Basis for Expansionary Policy of K. Takahashi," *The Annals of the Chuo Institute of Economic Research*, 1981, 201–26.
- Shima, K., "Iwayuru Takahashi Zaisei ni Tsuite," *Kinyu Kenkyu*, 1983, 2, 83–124.
- Takahashi, Korekiyo, (1936a) *Keizairon*, Tokyo: Chikusa Shobo, 1936.
- \_\_\_\_\_, (1936b) *Kokusaku Unyo no Sho*, Tokyo: Toan Shoin, 1936.
- \_\_\_\_\_, (1936c) *Zuisoroku*, Tokyo: Chikura Shobo, 1936.
- Yamamura, Kozo, "Then Came the Great Depression," in H. Van Der Wee, ed., *The Great Depression Revisited*, The Hague: Martinus Nijhoff, 1972.
- Japan, Ministry of Finance, *Okura Daijin Kaikoroku*, Tokyo: Okurasho, 1977.
- Bank of Japan, *Economic Statistics of Japan*, various issues.
- Oriental Economist*, *Oriental Economist Yearbook*, various issues.

## RISK PERCEPTION AND MARKET PERFORMANCE<sup>†</sup>

### Financial Risk and the Burdens of Contracts

By HERMAN B. LEONARD AND RICHARD J. ZECKHAUSER\*

A young artist we know was recently approached by a mail-order marketing company affiliated with a major credit card. She was asked to develop designs for use on sets of high-quality china to be marketed nationally. The company initially offered to purchase the designs on a straight commission basis—our friend would receive a percentage of gross sales. She asked to be paid a fixed fee instead.

This request seems odd. It might indicate that she distrusted her artistic capacity for this task (she didn't), that she would slack off (she wouldn't), or that she distrusted their marketing skills (she did, but she hardly wanted to tell them). The company's royalty proposal provides a financial incentive for quality, but it conveys a mixed message. To induce a large investment of the artist's time, the company should have tried to persuade her that it was confident of selling many sets. The royalty offer suggests instead that the firm fears only a few sets will be sold, and wants to protect itself against excessive artist's fees per set.

The negotiation concluded with the company agreeing to pay an attractive fixed fee. This arrangement placed the marketing risk on the company, which is likely both to perceive it as smaller and to be better able to bear it no matter what its size. It also removed any incentives to undermarket that might have been created by a royalty agreement.

But the contract gives the artist no direct incentive to spend the extra time to make these designs superlative (such efforts would be repaid by attracting future business, however).

This contract clearly attends to financial risk. The effects of variance in the outcome are borne by the party better able to absorb them. But it is difficult to assess the risk-assignment aspects—how much our friend "paid" to get the insurance she bought, or for the negative signal about confidence she had to send to get it—because they are intertwined with a collection of other messages and complicated by other motivations. A basic tenet of economics is that joint gains from trade arise from differences between the parties (Leonard and James Sebenius, 1983). But the parties must simultaneously discover and exploit their differences; thus a contract becomes not merely a division of known gains, but a device for revealing them as well. Even relatively simple contracts are asked to carry a whole collection of informational and incentive burdens. Risk perceptions (Amos Tversky and Daniel Kahneman, 1974), risk preferences, and the chosen allocation of risk between the parties (Kenneth Arrow, 1971) are three elements of the collection; they may be difficult to see through the veil of the others.

#### I. Multiple Burdens

Contracts must carry multiple burdens when outcomes—and possibly inputs—are uncertain or unknown. Financial risk is recognized in the form of contracts: in what will be exchanged, under what contingencies. But this is not the only issue a given contract is designed to resolve, and often it is not the most important one. The form of contracts becomes particularly important when at least

<sup>†</sup>*Discussants:* George Akerlof, University of California-Berkeley; Kenneth J. Arrow, Stanford University; Gary S. Becker, University of Chicago.

\*John F. Kennedy School of Government, Harvard University, Cambridge, MA 02138. We are indebted to Kenneth Arrow, Gary Becker, and Howard Kunreuther for helpful comments, and to Nancy Jackson for editorial assistance.



one party is in doubt about outcomes because of uncertainty about the inputs that can or will be supplied by another.

In the traditional market paradigm, certainty and full information prevail, and the function of contracts is simply to divide the known outcome(s) efficiently. For example, wages must be set so that conditions of competitive equilibrium in labor markets are satisfied. Contracts simply specify exactly what will be exchanged, governed by what contingencies, and at precisely what price. Modern economics, in contrast, has examined a variety of forms of ignorance and the modifications of contracts that may arise to cope with them. Ignorance about random influences on *outcomes* (i.e., pure uncertainty about the future values of random variables) is in many ways the simplest form of missing information. Actual outcomes vary around their expected values; financial theory examines the consequences of that variance in a world where individuals have different perceptions of risk and different aversions to it. It prescribes modifications of contracts that efficiently divide both the expected outcome and the variance, and capture all relevant information possessed by at least a few participants in the market. If these were our only forms of ignorance, and if no information were closely held, financial markets could readily deal with risk, and we would be able to discern that they were doing so.

Two additional major forms of ignorance—about *characteristics* and about *performance*—have important consequences for contracting, however. Signaling theory is concerned with how we use observed purchases, which may include contract provisions offered, to reveal what cannot otherwise be discovered at such low cost (Michael Spence, 1974). Often we can arrange contracts or set conditions before entering contracts that will successfully distinguish people with materially different but unobservable characteristics, even when some of them would prefer to mask the differences. Agency theory starts from the proposition that principals cannot costlessly monitor the efforts or performance of agents they retain (Michael Jensen and William Meckling, 1976; John Pratt and Zeckhauser, 1985). Accord-

ingly, contracts must be concerned with providing motivation—that is, with enforcement (or at least support) of the intended bargain.

Simple contracts cannot always address all of these concerns simultaneously. The needs are at least distinct, and they are often in conflict. By analogy to a lesson of Jan Tinbergen, a few contract parameters may not be enough to address the concerns separately. Moreover, interactions of contract burdens can result in unexpected inefficiencies. For example, signals may be costly to obtain, but once purchased they do not typically affect marginal payoffs, and hence do not distort incentives. But when signaling is part of a complex package that must provide motivation and distribute risk, the resulting imperfect contract often does involve marginal distortions. Were there no risk aversion, such difficulties could be avoided, by merely letting the agent reap all the benefits from the outcome, that is, by selling him the company. If more than one agent is contributing effort, even this scheme will not work.

## II. Worker Selection

These conflicts can arise even when there is no uncertainty about future random occurrences. Consider a firm that wants to hire the best available salesperson—the one who will sell most. Potential workers know exactly how much they would sell, but the firm does not. Each worker has the same reservation wage. Effort is not involved, just human capital (Gary Becker, 1964). The contract must both identify the best salesperson and divide the value of production between the firm and the worker. Clearly, no simple salary offer can distinguish among workers (either all or none would want the sales job, depending on whether the sales salary exceeded the reservation wage). A commission contract does provide this distinction. If the firm knows how good the best salesperson is, it can simply offer a commission rate low enough so that only the salesperson with greatest ability can earn more than the reservation wage. If the firm does not know the maximum ability in the population, it can still arrange the appropriate selection. For example, it can offer the job at auction,

hiring whoever agrees to accept the lowest commission rate.<sup>1</sup>

If we ask more of this contract, it too becomes insufficient. Suppose, for example, that actual sales are distributed around expected values known to workers, and that potential salespeople are risk averse while the firm is not. Then potential sales workers' desires to purchase insurance will conflict directly with the need to reveal which among them is the most able. The equilibrium must involve some use of royalties as a device to distinguish ability levels. But the participant selected through this sieve would prefer that the bargain be in the form of salary only. Similarly, a contract like this may partially reveal and accommodate differences in perceptions of risk, in work disincentives (and therefore prospective effort levels), and other features that both parties cannot perfectly observe—but only in exceptional circumstances will it be possible for a reasonably simple contract to achieve a fully efficient result.

A second example shows the interaction of risk aversion with other contract conditions more fully. Suppose we have three types of potential sales workers, each with reservation wage 100, with risk aversion and human capital (indexed by a normal distribution on output) as indicated:

	Salesperson Type		
	1	2	3
Potential Sales Profit			
Mean	100	110	120
Standard Deviation	10	10	30
Risk Aversion	none	severe	severe

What form of contract will both select the right salesperson for the job and distribute risk efficiently between whoever is selected

and the risk-neutral firm? One strategy would be to offer the firm for sale to the highest bidder, at the price quoted by the second highest bidder. This has the standard incentive compatibility of any Theodore Groves (1973)/Edward Clarke (1971) scheme, but honest elicitation is not enough in this case. Given sufficient risk aversion on the part of bidders 2 and 3, the highest bidder will be 1—the least able of the lot. We can elicit signals about who is best by selling a stake in the firm—for example, we might set a commission rate of 10 percent and give the position to whoever will accept the lowest salary. But this may lead to the selection of bidder 2 over bidder 3.

A less common contract does manage to select the right bidder. Suppose we offer a salary of \$99 and announce that we will also offer profit sharing to whoever bids the lowest share (we refer to such a contract as a "sliver," since a small fractional equity stake in the enterprise has been transferred to the agent). To maintain incentive compatibility, the sharing rate will be whatever is bid by the second lowest bidder. The fact that the fraction is small insures that the effect of risk aversion is removed, and the sliver system selects the most productive salesperson.<sup>2</sup>

It may not be practical to offer a profit-sharing plan of the sliver type—the commission rates may be so low as to seem silly, for example. Another contract, which we have not seen used, will work instead. Let the firm announce that it will pay \$99 in salary plus a \$2 bonus (an amount small enough not to incur risk-aversion losses) to whomever it hires if he or she meets a sales quota. The potential worker who volunteers to accept the highest quota will be hired, with the quota set at the level offered by the second highest bidder, for incentive compatibility.

<sup>1</sup>In Appendix I (available on request), we spell out a more elaborate example in which neither a simple salary offer nor a commission offer successfully separates workers by ability in a competitive labor market setting. Only a combination of salary and commission can support an efficient equilibrium. Even this more complex contract structure is unable to provide appropriate separation when we add to the contract the further burden of allocating risk efficiently.

<sup>2</sup>This contract breaks down as a selection device, however, if different bidders have different reservation wages. Suppose, for example, that bidder 3 has a reservation wage of 105. Given an excess productivity of 10, he or she is still the best candidate for the job. But now a contract of \$99 (or any constant) plus a small equity stake will not be enough to separate bidder 3 from bidder 2.

The contract is a small bet, so risk aversion will not play a role. This contract based on a bonus for meeting a negotiated performance standard thus permits identification of the most able salesperson and appropriately allocates risk.

Both the need for a mechanism for revealing personal characteristics and the need to build in performance incentives are often in direct conflict with the parties' preferences for dividing risk. People generally know more about their own characteristics than others know about them, and they often have a better assessment of the risks involved (if any). Consider the entrepreneur-inventor approaching venture capitalists. The inventor may be unduly optimistic about the market's reception of new products, but almost surely has a better technical understanding of the potential—and the problems—of the proposed product. This more accurate assessment comes bundled with a conflict of interest about revealing it truthfully. Any contract with a supplier of capital must bear the burden of revelation as well as division. The most natural way to do this would be to offer contracts that will only be accepted by inventors with high subjective probabilities that their inventions will succeed—that is, contracts that will pay off handsomely only if the product is a considerable commercial success. But these are exactly the contracts that impose the greatest financial risk on the entrepreneur, the party to whom it is probably most costly.

If we resolve this conflict in favor of appropriate risk spreading—pay a fixed fee for the patent and put the inventor on salary—the contract can no longer reveal honest risk assessments by the most knowledgeable party. If we offer a strongly revealing contract—a small salary with a large royalty rate on sales beyond the breakeven point—the contract cannot allocate risk bearing efficiently. Moreover, the magnitude of the risk burden will vary directly with the incentive for the entrepreneur to work—when the acceptance of risk is being used as a signaling device, it often distorts work incentives at the margin. (This problem is not unique to signaling through risk, but in many situations the signal, once purchased, does not

alter marginal incentives.) Neither extreme on this contracting spectrum is particularly attractive. In practice, we are likely to observe some form of compromise—which means that full optimality is sacrificed and that it may be hard to discern just how (and how much) the market arrangement has recognized financial risk.

### III. Responses to Risk Sharing

Providing performance incentives in contracts also frequently militates against appropriate risk spreading. Consider the problem of rewarding managers. For a collection of well-understood reasons, shareholders want the firm not to be risk averse. But managers face career risk within the firm and are themselves understandably risk averse. As a result they may be unwilling to undertake sufficient risk. We often observe incentive contracts for managers (stock options, for example) that at least partially balance this (inefficient) caution. They are valuable only when the company does quite well—an event made more probable if the managers take greater risks. (A standard result of options theory is that ownership of options makes one risk preferring, or less risk averse.) But this performance incentive, deliberately designed to offset the risk aversion of managers, must proceed precisely by imposing greater risk on them. Indeed, studies of compensation for high-level executives have found very high levels of risk borne by the agents—evidently, the advantages from performance inducements and the selection gains from attracting risk takers compensate the losses from poor risk spreading. This makes it difficult to uncover the elements of these contracts that recognize and deal with financial risk—they are deeply intertwined with other burdens the contracts are carrying.

The need for incentives to balance the nonlinearity of rewards with performance can run in the other direction. Western legal tradition confers limited liability on shareholders, producing an asymmetry in payoffs around the point of bankruptcy. When a firm is near bankruptcy, it is protected from the full adverse consequences of any gamble, and this may induce risk-preferring behavior.

Indeed, it results in a fundamental managerial principle: "when your back is against the wall, roll the dice." Under these circumstances the firm may engage in gambles for which the risk-adjusted rate of return is below the risk-free rate, that is, it may accept lotteries involving a mean-sacrificing spread (*MSS*) of outcomes.<sup>3</sup>

#### IV. Robustness and Optimality

Since the many burdens contracts are asked to carry are often in conflict, full optimality is difficult to achieve.<sup>4</sup> But why do we not more often see more complex contracts, tailored to address a wider set of concerns? One answer, often advanced, is that customizing contracts is expensive. We propose an additional reason: like creatures too carefully adapted to a particular ecological niche, highly specialized contracts are vulnerable to small changes in conditions. For example, if we must address only risk aversion and selection on the basis of ability through a contract offer, a contract offering a high salary and a small equity stake in the outcome will suffice (the small equity stake attracts the more able; the fact that it is small means that it does not impose much risk and therefore involves no real loss due to risk aversion). If we need not worry about risk aversion, but must provide a marginal incentive for effort along with a selective mechanism to identify the more able, then a large equity stake in the outcome will generally be appropriate. In either case, a simple contract, specialized to address the precise conditions defining the contracting context, suffices.

<sup>3</sup>Given a risk-free rate and a market premium for risk, a gamble is a mean-sacrificing spread if it involves below-market compensation for risk bearing. The expected return may be positive; it is still a sacrifice if the mean rises by less than we would expect, given the risk. Firms near bankruptcy might be induced to accept even gambles that involve an absolute reduction in mean in return for a sufficiently large spread. The propensity to engage in *MSS* can be offset by instruments such as bonds convertible to stock at the bondholders' option.

<sup>4</sup>In Appendix II (available from the authors), we describe a scheme for categorizing contracting situations and contract types, and comment on whether optimality can be achieved under various combinations of them.

Unfortunately, a small departure from the specified underlying conditions will make either contract seriously deficient. If agents are risk averse and we must both select the best workers and encourage effort at the margin, neither a small nor a large equity stake will work well—one provides too little incentive for effort, the other too much risk. Adding even a small payoff from effort to a contract context that originally involved only selection and risk aversion could dramatically alter the contract stipulations required to achieve optimality. The carefully arranged near optimality of a specialized contract crafted to address a given set of conditions can often be seriously upset by even a small dose of some other condition.<sup>5</sup> Thus familiar contracts, frequently simple in form, may not only be only cheaper to arrange, but may also be more robust. That is, they may achieve reasonable, though imperfect, results across a wider range of contracting conditions. This may be an important part of their attraction. Simplicity and robustness are, of course, not synonymous; we are merely observing that contracts designed to address fewer special features seem less likely to be blindsided by others.

The attractiveness of simple familiar contracts, in turn, makes it difficult to identify and measure the extent to which financial contracts appropriately recognize risk. Our ignorance is much more profound than simple market uncertainty. Contracts must specify divisions of benefits, but they must also

<sup>5</sup>In many contracts, the degree of optimality obtained is a smooth function of the underlying conditions, and the characteristics of the optimal contract vary continuously as a function of the burdens being carried. For example, a small change in the risk aversion of the agent in a contract balancing performance incentives against risk spreading would call for only a small adjustment in contract terms. By contrast, when selection is involved a small alteration in conditions can lead to a switch in which candidate gets the job, and a consequent dramatic shift in the degree of optimality obtained. A major change in the form of the contract may be required to maintain optimality. For example, our negotiated quota arrangement can fail if the three salesmen have different reservation wages—and even a small change in reservation wages can take us across the boundary.

provide motivation even as they rely on it, and provide information about personal characteristics even as they make use of it. Because contracts must simultaneously address all of these needs, the result is either complicated arrangements or, more typically, uncomfortable compromises within simple contracts. This makes it difficult to see how any particular issue is being dealt with. We see only the congealed Gordian knot; the separate strands can no longer be distinguished.

#### REFERENCES

- Akerlof, George, "The Market for Lemons," *Quarterly Journal of Economics*, August 1970, 84, 488-500.
- Arrow, Kenneth, *Essays in the Theory of Risk Bearing*, Chicago: Markam, 1971.
- Becker, Gary, *Human Capital*, New York: Columbia University Press, 1964.
- Clarke, Edward H., "Multipart Pricing of Public Goods," *Public Choice*, Fall 1971, 11, 17-33.
- Groves, Theodore, "Incentives in Teams," *Econometrica*, July 1973, 41, 617-31.
- Jensen, Michael and Meckling, William, "Theory of the Firm, Managerial Behavior, Agency Costs, and Ownership Structure," *Journal of Financial Economics*, October 1976, 3, 305-600.
- Leonard, Herman B. and Sebenius, James K., "Differences and Gains from Trade," mimeo., Kennedy School of Government, 1983.
- Pratt, John W. and Zeckhauser, Richard J., *Principals and Agents: The Structure of Business*, Boston: Harvard Business School Press, 1985.
- Spence, Michael, *Market Signalling*, Cambridge: Harvard University Press, 1974.
- Tversky, Amos, and Kahneman, Daniel, "Judgment under Uncertainty: Heuristics and Biases," *Science*, September 1974, 185, 1124-31.

# Are Individuals Bayesian Decision Makers?

By W. KIP VISCUSI\*

There has been increasing interest in whether normative models of individual choice under uncertainty accord with actual behavior. These concerns have been much greater than in other economic contexts because of the particularly severe demands such decisions place on the rationality of the decision maker. The limitations of these decisions have widespread consequences, as they provide the rationale for many governmental efforts to regulate the risks people face. Here I explore the issues raised by a Bayesian decision framework, focusing particularly on my analyses of worker and consumer behavior.

## I. Risk Perceptions

Ideally, individuals should fully understand the risks they face before making decisions with probabilistic outcomes. In most instances, extensive experimental evidence is not available, so that individuals must rely on their subjective probabilistic judgments. Such assessments will clearly not always be accurate and may be systematically biased as well. Precise analysis of the nature and extent of such biases is impeded by the paucity of data on individuals' probability assessments and the actual risks that they face.

My analysis (1979) of worker risk perceptions focused on survey data for which I linked an objective risk index (the BLS injury rate for the worker's industry) and a measure of the worker's subjective risk perceptions—a dummy variable for whether or not the worker's job exposed him or her to dangerous or unhealthy conditions. The expected positive correlation was observed, but such evidence can only be suggestive because the workers did not scale the risks in probabilistic terms.

To refine this evidence, in my article with Charles O'Connor (1984) we presented over 300 chemical workers a linear scale that they would use to rate the hazards of their jobs. This scale was constructed in a manner that made it possible for us to compare workers' responses to objective measures of the chemical industry risk. In particular, each rating could be converted into an equivalent level for the BLS injury and illness rate. Overall, workers' subjective risk assessments were above the reported injury and illness rate for the chemical industry, which one would expect since health risks are notoriously under-reported. What was particularly noteworthy is that once the long-term chemical hazards were excluded from consideration (for a subsample that was told that they would be working with sodium bicarbonate instead of their present chemicals), the subjective risk perceptions were identically equal to the published accident rates. Although one would be hard-pressed to claim that such a fortuitous result implies that all job risk perceptions are unbiased, there does appear to be a strong correspondence between actual and perceived risks for a major class of risks that people face.

When asked to rate their job risks using a linear scale or when asked about whether or not their jobs pose a hazard, most respondents give plausible risk-perception assessments. These assessments are much more accurate than the responses in studies that frame the risk perception issue in relative terms (for example, whether or not the respondent believes he or she is an above-average risk driver), where systematic optimism has been observed. Some observed biases in past studies may be due to the manner in which the risk-perception question is framed, rather than any underlying shortcoming in individual behavior.

It is well known, however, that individuals have particular difficulty in thinking about low probability events. An especially influential study is that of Sarah Lichtenstein et al.

\*Professor and Director, Center for the Study of Business Regulation, Fuqua School of Business, Duke University, Durham, NC 27706.

(1978), who explored fatality risks, ranging from tornadoes to strokes and homicides. The pattern they observed was that individuals overassessed small risks of death and underassessed large risks.

Although these biases in risk perception are widely cited as evidence of irrationality, in my forthcoming article I show that such a pattern is exactly what one would expect from a Bayesian learning process. Let  $p_i^*$  be the probability of death from accident category  $i$ ,  $s_i$  be the actual risk from category  $i$ , and  $p$  be the prior risk assessment before knowing the category to which the accident belongs. If individuals behave rationally and their probability assessments are characterized by a *beta* distribution, one can show that

$$(1) \quad p_i^* = (p + \psi_i s_i) / (1 + \psi_i),$$

where  $\psi_i$  is the relative informational content associated with category  $i$  compared with the individual's prior. More specifically,  $\psi_i$  represents the equivalent number of Bernoulli trials that the individual acts as if he (or she) has drawn for category  $i$  accidents divided by the number of trials he acts as if he has observed when forming his prior  $p$ .

Individuals' risk perceptions should be a weighted average of the true risk and their prior, where the weight depends on how much information they have about category  $i$ . The value of the prior risk assessment  $p$  was not ascertained in the survey so I used two different proxies for this risk in separate equations. The first was an Akerlof "lemons" model measure—the average of all of the risks in the sample. The second proxy used was the reference point that each respondent was given in the survey before assessing the risks (either motor vehicle deaths or electrocutions).

The subsequent empirical estimates were consistent with the linear relationship specified in equation (1). For small risks, individuals revised their prior beliefs downward, but not fully, leading to overestimation of the risk. Similarly, for large risks the prior is revised partially in the upward direction, leading to an underassessment of the risk. It

is particularly noteworthy that the relative weight  $\psi_i$  placed on the true risk level was not significantly correlated with the degree of risk so that there was no bias in the manner in which probabilistic beliefs are revised in the direction of the true risk.

Overall, the evidence suggests that individuals may have reasonably accurate perceptions of risks that have a fundamental impact on their welfare. Risk perceptions for other more remote hazards are less precise, and the observed biases are exactly what one would expect from a rational, Bayesian learning process. The inadequacies in risk perception also do not appear to be clearcut in either direction. The overestimation of small risks and underestimation of large risks represents a more complex type of market failure than is reflected in the usual economic models incorporating biases in risk perception, which typically assume that risks are underestimated.

## II. The Role of Learning

The cornerstone of the Bayesian approach is the learning process by which individuals update their risk perceptions. This learning process was implicitly involved in the formation of the risk perceptions discussed above. In my 1979 study, I analyzed the consistency of workers' risk perceptions with the possibility of on-the-job learning using cross-sectional data. Workers who had experienced a job injury or viewed other working conditions as being unpleasant were more likely to view their jobs as being dangerous, controlling for the industry risk level and related factors. One cannot be confident based on this evidence that workers do in fact learn, since the results may simply reflect the correlation of high initial risk assessments with risky job attributes.

To explore the evolution of workers' risk judgments, O'Connor and I undertook an experiment with chemical workers at four plants. Each worker was shown a label for a chemical (either sodium bicarbonate, chloroacetophenone, asbestos, or TNT) and was told that this chemical would replace the chemicals on his present job. Workers' risk perceptions responded in the expected man-

ner, falling for sodium bicarbonate and rising for the other three chemicals. Since the true job-specific risks posed by these substances and the worker's other job risks is not fully understood, the most that could be concluded is that the prior probability assessments were revised in the correct direction.

Based on the worker responses, it was possible to estimate the key parameters in equation (1) that are associated with the label's impact—the risk  $s_i$  implied by the label and its relative informational content  $\psi_i$ . Except for sodium bicarbonate, the implied risks  $s_i$  did not differ greatly. There was, however, a substantial difference in the relative informational content  $\psi_i$  of the hazard warnings. The unfamiliar chemical chloroacetophenone had a  $\psi_i$  value of 1.3, implying an informational content just above that of the worker's prior, whereas TNT had a dominant  $\psi_i$  value of 31.4.

These results suggest that people can process risk information in the expected direction, but that it is the informational content of the message, not simply the associated risk level, that is instrumental. The ineffectiveness of informational campaigns to promote seatbelt use and to discourage cigarette smoking are not unexpected, since the new information contained in such ads is not great. These efforts might be viewed more accurately as being policies of exhortation rather than information.

While available evidence suggests that individual learning about risks can often play an important economic role, this learning process may not be ideal. The critical reviews by Amos Tversky and Daniel Kahneman (1974), Kenneth Arrow (1982), and Baruch Fischhoff and Ruth Beyth-Marom (1983) have identified a number of systematic shortcomings. Individuals tend to exaggerate the completeness of hypothesis sets, ignore the base-rate frequency of outcomes, and more generally fail to fully understand the laws of probability. Individuals may behave in the general spirit of Bayesian decision makers in the learning process, but this behavior does not conform identically with an optimal learning process. The degree to which the various shortcomings identified in lab-

oratory experiments affect market behavior involving risks has not yet been ascertained.

### III. Risk Perceptions and Individual Behavior

A number of studies have linked higher wages to job risks, consistent with Adam Smith's theory of compensating differentials. This is the most basic test of rational decisions involving uncertainty, and the supporting evidence is strong and quite diverse. Compensating differentials have been estimated for a large number of data sets using a variety of risk measures. It is particularly noteworthy that the observed premiums are roughly similar whether the risk variable is an objective hazard measure (for example, the industry death rate), or a measure of subjective risk perceptions (see my 1979 study and my article with O'Connor). Since it is the subjective perceptions that are instrumental from an economic standpoint, these results suggest that the fundamental behavioral assumption of the theory is satisfied.

Although the risk level is the only feature of the job risk that is of consequence in a single-period model, in a multiperiod model in which there is the possibility of terminating the job either through a job change or one's death, I have shown that the precision of the risk judgments is an addition concern. The underlying rationale is that in this class of two-armed bandit models, loose prior beliefs are preferred because they offer the potential for greater gains from experimentation with the uncertain job. As predicted, chemical labels associated with higher  $\psi_i$  values lead to higher worker reservation wages, as do labels with higher  $s_i$  values. Both the risk level and its precision affect a lottery's attractiveness if one is incurring a sequence of such lotteries that may be terminated conditional on an unfavorable outcome.

Although these results are consistent with optimal behavior in uncertain contexts with learning, not all observed risk-dollar tradeoffs imply that decisions are accurate. In a recent study of consumer attitudes toward low probability events (on the order of 1/100,000 risk annually), Wesley Magat and I (1984) ascertained consumers' valuations of different health outcomes. The results suggested



implausibly large risk-dollar tradeoffs. For example, there was an implied externality value to society at large of roughly \$200,000 to prevent a hand burn from drain cleaner that would be temporary, but severe enough to require medical treatment. Individuals clearly have difficulty making decisions involving low probability events, and in this instance there is evidence of excessive valuation of the risks. These biases in turn may lead to alarmist decisions and excessive governmental regulation. In other cases the low risk may be ignored, creating biases of the opposite nature.

Once learning is introduced as an element, individuals will continually reassess the appropriateness of the risks and its rewards in relation to their other opportunities. The tendency of individuals to experiment with activities posing dimly understood risks will be fostered by the structure of the statistical decision problem. Individuals will display a predilection for risky jobs and other lottery sequences associated with loose priors since these offer the greatest gains from experimentation.

As predicted by these two-armed bandit models, there is a significant relationship between job risks and worker decisions to switch jobs once significantly adverse information is acquired. Results for five data sets reported in my 1979 study indicate that job risks raise worker quit rates, boost quit intentions and job-search activities, and shorten paths of employment at the firm, controlling for health status and a variety of other factors. Indeed, job risks may account for as much as one-third of all manufacturing quit rates.

In addition to the positive effect of the risk level on quitting, the aforementioned work on chemical labeling produced a positive influence of the relative precision  $\psi_i$  of the risk information on quit behavior. This impact is also consistent with an optimal experimentation process since more precisely understood risks are less attractive because of the diminished value of the experimentation process associated with them. The overall job choice process is consistent with a model in which individuals start jobs with imperfect information, revise these beliefs in Bayesian

fashion based on their on-the-job experiences, and alter their job choice if this information is sufficiently adverse.

Consumers likewise respond to risk information in an adaptive manner. In our study of consumer product labeling, Magat and I found that labels including risk warnings increased the frequency of consumer precautions by up to 33 percent, as compared with labels without warnings. These results may understate the role of learning to the extent that consumers' prior beliefs have been conditioned by past knowledge of the product. For example, even without a hazard warning, more than half of all consumers would store drain cleaner in a location to which children did not have easy access.

Overall, individuals do not possess perfect information about the risks they face, but they do have opportunities to revise these beliefs based on their experiences. The observed behavior patterns are consistent with the principal predictions of a Bayesian learning process and subsequent adaptive behavior.

Although this behavior is broadly consistent with a Bayesian framework, these decisions do not always coincide with optimal behavior. As with other optimizing models in economics, Bayesian decision models represents an often powerful tool, but also a tool that may not accurately reflect how decisions are made. The expected utility hypothesis that is central to these models has long been questioned. In some contexts, inconsistencies in individual choices have been observed. There also appears to be asymmetric treatment of gains and losses, as well as special attention paid to certain outcomes. Moreover, in an actual market context in which one would have expected risk-averse consumers to purchase heavily subsidized flood and earthquake insurance, Howard Kunreuther et al. (1978) have shown that they failed to do so. As a result, individuals may respond in a manner that is broadly consistent with Bayesian decision theory, but the normative guidelines of that theory may not always be met.

Nevertheless, Bayesian models remain a useful optimizing framework for analyzing economic behavior. In some cases, the ex-

isting biases in behavior may be predicted by proper application of the Bayesian model. In others, there may be shortcomings in the manner in which individuals make decisions.

The implications of these inadequacies for the nature of the market failure are not always clear-cut. Risks may be ignored, leading to a supra-optimal level of risk, or they may be over-assessed, as shown in studies of small fatality risks. The nature of the market failure is likely to be more complex than is captured in standard models of imperfect information. There may be either inadequate or excessive attention to risks, depending on the particular context. Much remains to be learned about the shortcomings of individual decisions, the magnitude of these shortcomings, and their implications for market performance.

#### REFERENCES

- Arrow, Kenneth J., "Risk Perception in Psychology and Economics," *Economic Inquiry*, January 1982, 20, 1-9.
- Fischhoff, Baruch and Beyth-Marom, Ruth, "Hypothesis Evaluation from a Bayesian Perspective," *Psychological Review*, 1983, 90, 239-60.
- Kunreuther et al., Howard, *Disaster Insurance Protection: Public Policy Lessons*, New York: Wiley & Sons, 1978.
- Lichtenstein et al., Sarah, "Judged Frequency of Lethal Events," *Journal of Experimental Psychology*, 1978, 4, 557-78.
- Tversky, Amos and Kahneman, Daniel, "Judgment under Uncertainty: Heuristics and Biases," *Science*, 1974, 185, 1124-31.
- Viscusi, W. Kip, "A Bayesian Perspective on Biases in Risk Perception," *Economics Letters*, forthcoming.
- \_\_\_\_\_, *Employment Hazards: An Investigation of Market Performance*, Cambridge: Harvard University Press, 1979.
- \_\_\_\_\_, and Magat, Wesley, "Analysis of Economic Benefits of Improved Information," Report to U.S. Environmental Protection Agency, 1984.
- \_\_\_\_\_, and O'Connor, Charles, "Adaptive Responses to Chemical Labeling: Are Workers Bayesian Decision Makers?," *American Economic Review*, December 1984, 74, 942-56.

# Ambiguity and Insurance Decisions

By ROBIN M. HOGARTH AND HOWARD KUNREUTHER\*

Imagine that you are in the insurance business and are preparing to quote premiums for two risks that can be characterized as follows:

In Situation *A* there is a potential loss of \$100,000 and your best estimate of this loss occurring within the period covered by the insurance is .01. Moreover, your estimate of the probability is based on the opinions of several experts all of whom agree on the .01 figure.

In Situation *B* there is also a potential loss of \$100,000. However, whereas your best estimate of the probability of this loss occurring within the contract period is .01, this figure is a compromise reached after considering the widely different estimates furnished by several experts.

When comparing these two scenarios, many persons deem Situation *B* to be inherently more risky than Situation *A*, such that the premium required to insure the latter would be larger than for the former. However, it is important to note that this behavior contradicts the expected utility model which economists have typically used to analyze insurance decision making (see Isaac Ehrlich and Gary Becker, 1972). The reason is that this model does not distinguish between cases where people do or do not experience uncertainty or ambiguity about their estimates. However, as demonstrated by Daniel Ellsberg (1961) and others (for example, see Selwyn Becker and Fred Brownson, 1964) in experimental settings, ambiguity or "uncertainty about one's uncertainty" does

affect choice. On the other hand, if one believes that ambiguity affects behavior in markets, one needs a model that predicts how such effects occur and when they are important. Our purpose here is to explore the predictions of such a model with respect to the market for insurance.

First we sketch a psychological model of probabilistic judgment under ambiguity developed by Hillel Einhorn and Hogarth (forthcoming). This is then used to predict how both consumers (buyers) and insurance firms (sellers) are likely to react toward differing degrees of ambiguity. A brief summary of the results of experiments designed to test these predictions are presented. Finally, we discuss these results in relation to real world phenomena as well as considering possibilities for future empirical research. We particularly note the importance of developing and testing precise, falsifiable models to complement or challenge implications of the expected utility model since naturally occurring data frequently lack the power to provide stringent tests of the latter.

## I. Ambiguity and Choice

A model of how people assess probabilities of events in ambiguous circumstances has recently been proposed and experimentally tested by Einhorn and Hogarth. The model is based on the following three principles.

1) People are first assumed to anchor on an initial estimate of the probability. Let  $p$  represent the anchor and note that it may be based on past experience, suggested by an analogous situation, or even the figure provided by an expert. The anchor is then adjusted by imagining or mentally simulating other values that the probability could take.

2) The greater the degree of ambiguity experienced, the more alternative values of the probability are simulated and the larger the weight given to such values in the final

\*Graduate School of Business, University of Chicago, Chicago, IL 60637, and Wharton School, University of Pennsylvania, Philadelphia, PA 19104, respectively. We are grateful to Colin Camerer, Hillel Einhorn, Jack Hershey, and Paul Kleindorfer for helpful comments on an earlier version. This research is partially supported by NSF grant SES-8312123 and a contract from the Office of Naval Research.

assessment. Thus, when experts disagree on a probability estimate, people are assumed to imagine more alternative values compared to situations where the experts agree.

3) The weight given in imagination to alternative values of the probability that are greater or smaller than the anchor  $p$  depends on the individual's attitude toward ambiguity in the particular situation.

Let the adjustment to the anchor be represented by  $k$  so that the assessment of the ambiguous probability, denoted  $S(p)$ , is given by

$$(1) \quad S(p) = p + k.$$

To allow for the effects of ambiguity, Einhorn and Hogarth decompose  $k$  into two parts that capture forces favoring positive and negative adjustments, respectively. The positive force reflects the weight given to possible values of the probability above the anchor and is taken to be proportional to  $(1-p)$ ; the negative force reflects the weight given to values below the anchor and is proportional to  $p$ . In both cases, the constant of proportionality is a parameter  $\theta$  that represents the amount of perceived ambiguity in the situation ( $0 \leq \theta \leq 1$ ). In other words, the effect of possible values of the probability above the anchor are modeled by  $\theta(1-p)$ , of those below by  $\theta p$ , and  $k$  is the net effect of positive and negative adjustments from the anchor. However, to account for the fact that values above and below the anchor may be differentially weighted in imagination,  $\theta p$  is adjusted to the form  $\theta p^\beta$  where  $\beta (\beta \geq 0)$  represents the person's attitude toward ambiguity in the circumstances. That is, when  $\beta = 1$ , equal weight is given to imaginary values above and below the anchor; when  $\beta > 1$ , more weight is given to larger values; and for  $\beta < 1$ , more weight is given to smaller values. This leads to the model

$$(2) \quad S(p) = p + \theta[(1-p) - p^\beta].$$

Note that in this model  $\theta$  (i.e., perceived ambiguity) determines the amount of the adjustment, whereas  $\beta$  in conjunction with the level of  $p$  determines its sign. Thus, when  $p$  is low, the adjustment will tend to be

positive; however, as  $p$  increases, the adjustment will become negative. Moreover, the point at which the adjustment starts to become negative depends on  $\beta$ . That is, when  $\beta = 1$ , the "cross-over" point is at  $p = .5$ ; for  $\beta < 1$  this occurs when  $p < .5$ ; and for  $\beta > 1$  when  $p > .5$ .

Now assume that the price a consumer is prepared to pay for a given level of insurance coverage is a monotonically increasing function of  $S(p)$ . Furthermore, let  $C(p)$  denote the premium a consumer is prepared to pay for this coverage when the probability of the loss can be precisely estimated, and  $CA(p)$  that when the consumer is ambiguous. Assume further that when facing the ambiguous probability of a loss,  $\beta$  is not a small value (see discussion above). These assumptions imply the following predictions:

1) when  $p$  is small, the ratio  $[CA(p)/C(p)]$  will be greater than one indicating "ambiguity aversion"; 2) as  $p$  increases,  $[CA(p)/C(p)]$  will decrease and eventually become smaller than one thereby indicating "ambiguity seeking." Denoting the premiums that firms are prepared to charge with precise probabilities as  $F(p)$ , and  $FA(p)$  for premiums under ambiguity, similar predictions can be made for firms—but with one exception. The exception is that one would never expect firms to charge less than expected value. Therefore, for firms, the prediction is that the ratio  $[FA(p)/F(p)] > 1$  for small probabilities but that this ratio will decrease as  $p$  increases.

## II. Experimental Evidence

We have collected experimental data testing the above predictions in a series of questionnaire experiments involving some 500 individuals who were participating in undergraduate, graduate and executive program courses at the University of Chicago and the Wharton School. Most of these subjects had been exposed to courses in economics and statistics and thus were relatively sophisticated concerning insurance. Since all the experiments produced similar findings, we only report here some results from two experiments (for full details, see our 1984 paper).

In the first experiment, 113 University of Chicago MBA students responded to questionnaires administered in a class on decision making. The experimental stimulus concerned the owner of a small business who was considering insuring against a potential loss of \$100,000 from claims due to a possible defective product. Subjects played the role of either the owner of the business (buyer) or the head of the department in a large insurance company (seller) who was responsible for setting premiums. Each subject responded to both the ambiguous and nonambiguous versions of the stimulus (see below) at one level of the anchor  $p$  in an experimental session during which responses were also made to questions that were not related to this study. In all, four levels of  $p$  were investigated,  $p = .01, .35, .60$ , and  $.90$ . In both the ambiguous and nonambiguous cases a specific probability level was provided in the stimulus (for example,  $.01$ ). However, a comment was also added as to whether one could "feel confident" (non-ambiguous case) or "experience considerable uncertainty" (ambiguous case) concerning the estimate. Uniformity of perceptions of ambiguity was controlled by describing the situations by the same words in both the consumer and firm versions.

Table 1 reports median prices of the minimum premiums firms were prepared to charge and the maximum premiums consumers were prepared to pay, together with the ratios of ambiguous to nonambiguous prices at each probability level. Note that in the nonambiguous condition, the median prices are close to expected value for both firms and consumers. However, there are marked differences in the ambiguous case. For firms, premiums in the ambiguous condition exceed those in the nonambiguous case at all levels except  $p = .90$ . Consumers, on the other hand, are only prepared to pay more in ambiguous circumstances when the probability of loss is low (i.e.,  $p = .01$ ). At  $p = .35$ , consumers' maximum prices are the same for both levels of ambiguity and for larger probabilities ( $p = .65, .90$ ) they exhibit ambiguity preference in that maximum prices under ambiguity are lower than when there is no ambiguity. These results are consistent

TABLE 1—MEDIAN PRICES FOR INSURANCE<sup>a</sup>

Probability Levels	Ambiguous (1)	Nonambiguous (2)	Ratio
			Col. (1)/ Col. (2)
Consumers' Willingness To Pay			
.01	\$1,500	\$1,000	1.50
.35	\$35,000	\$35,000	1.00
.65	\$45,000	\$65,000	.69
.90	\$60,000	\$82,500	.73
Firms' Supply Price			
.01	\$2,500	\$1,000	2.50
.35	\$52,500	\$37,500	1.40
.65	\$70,000	\$65,000	1.08
.90	\$90,000	\$90,000	1.00

<sup>a</sup> Loss = \$100,000; a complete statistical analysis of these data supporting the statements made in the text is to be found in our earlier paper.

with the predictions of the Einhorn-Hogarth ambiguity model.

Whereas the above analysis is useful for testing the predictions of the ambiguity model across the range of  $p$ , it is less illuminating from a market perspective. The real issue here is whether willingness to trade on the part of firms and consumers is also affected by ambiguity. This issue was tested in a second experiment, also involving University of Chicago MBA students, who were again assigned either the roles of consumers or firms. Using the same scenario as the previous experiment, subjects were required to respond by stating whether they would trade ("Yes" or "No") at a given price. Having answered this question, subjects turned a page in their experimental booklets and were asked the same question with respect to a different price. To simulate trading conditions, the second price for consumers was lower than the first, whereas the reverse order was used for firms. Since the market for insurance typically covers situations where probabilities of losses are small, we only report here data for  $p = .01$ . The prices quoted were \$1,500 and \$3,000 and the results of this experiment are presented in Table 2. The results of this experiment are consistent with our earlier findings. Consumers are prepared to buy insurance at premiums well in excess of expected value for a low probability event ( $p = .01$ ). Moreover, there

TABLE 2—PERCENTAGES OF SUBJECTS PREPARED TO TRADE AT GIVEN PRICES<sup>a</sup>

	Prices Offered		n
	\$1,500	\$3,000	
Consumers:			
Ambiguous	87	83 <sup>b</sup>	23
Nonambiguous	77	50	22
Firms:			
Ambiguous	16	36	25
Nonambiguous	67 <sup>b</sup>	87 <sup>b</sup>	15

<sup>a</sup>Potential loss = \$100,000;  $p = .01$ .

<sup>b</sup>Differences between ambiguous and nonambiguous significant ( $p < .01$ ).

is greater willingness to insure at high prices when ambiguity is present than when it is not (83 vs. 50 percent). Some firms are also willing to provide insurance for such events. In addition, as the price increases, more firms are prepared to do business in both the ambiguous and nonambiguous cases. However, there is a marked difference in willingness to insure at the stated prices between the ambiguous and nonambiguous cases, that is, 16 vs. 67 percent at \$1,500, and 36 vs. 87 percent at \$3,000.

### III. Implications for Market Behavior

The effect of ambiguity can be precisely isolated in experimental situations, but this becomes far more difficult in natural settings. Insurance markets are rich contexts for studying ambiguity because insurers can always specify a maximum amount of coverage, but may be highly uncertain as to the probability that they will experience a loss of this magnitude. This effect of ambiguity combined with problems of adverse selection, moral hazard, and fear of bankruptcy may result in limited insurance markets. Consider the following two illustrations.

Less than 5 percent of California homeowners purchase earthquake coverage despite their belief that they will receive only limited aid from the federal government in the event of a large loss (Kunreuther et al., 1978). Insurance firms, on the other hand, charge high rates for coverage which may partially explain lack of consumer interest. Over the sixty-year period since coverage has

been offered in California, \$269 million in total premiums have been collected and only \$9 million in losses have been experienced (Arthur Atkisson and William Petak, 1981). These figures by themselves may not be unreasonable if there is a chance that a large earthquake would create huge losses. Furthermore the spector of bankruptcy could restrict this coverage to only large firms who would then have some monopoly power. The counterargument is that damage to residential homes is relatively minor even in a severe earthquake, so that all insurance firms should have an interest in providing coverage at more reasonable rates.

Special institutional arrangements have emerged for dealing with insurance against low probability high loss events. For example, the Price Anderson Act passed in 1957 currently offers \$160 million insurance protection by the private sector to cover nuclear accidents, with the government handling potentially large losses up to a maximum of \$560 million. The industry feels that it will be extremely difficult to increase coverage of private insurers on a voluntary basis due to the ambiguity associated with the probability of a catastrophic event (Alliance of American Insurers et al., 1979). Similarly, the passage of a federal-private flood insurance program in 1968 and the provision of political risk insurance, primarily by the Overseas Private Investment Corporation, are examples of programs for handling events where the probability of a given loss is ambiguous and where private insurers have provided only limited coverage.

Our principal reason for focusing on ambiguity is that it affects the way human judgment and choice deviates systematically from the rational models underlying much of economic analysis when both consumers (buyers) and firms (sellers) have the same information. However, instead of simply documenting another instance where the expected utility model fails to predict behavior (see Paul Schoemaker, 1982), we have specifically tested the implications of a model that does account for the effects of ambiguity. Note also that our model is similar *in spirit* to analyses of insurance markets that are based on the expected utility model. That is, start-

ing from some propositions about human choice behavior at the individual level, we proceed to make and test predictions at the aggregate market level. Moreover, we handle real phenomena that the expected utility model essentially assumes not to exist.

The impact of ambiguity on insurance markets seems to be particularly important for low probability events where there are limited opportunities for learning over time. For some events (for example, fire, life), consumers may have limited opportunities to learn while insurers have a wealth of experience from their claims files. Both parties may be uncertain as to the probabilities of other events (for example, earthquakes, chemical plant accidents etc.) due to limited past experiences and imperfect causal models. There is a need for future research to determine the extent of the ambiguity surrounding probabilities of different events consumers and insurers face, the impact this has on the premium structure and the extent of the market.

There is also a need for more theoretical and experimental work regarding the effects of "uncertainty about uncertainty" on behavior. Market-based experiments in the spirit of Charles Plott (1982) and Vernon Smith (1982) can further extend the questionnaire approach of this study to settings that mimic market conditions more accurately.

Finally, we recognize that ambiguity is but one of a number of phenomena such as regret (David Bell, 1982) and the context or "framing" of decisions (Amos Tversky and Daniel Kahneman, 1981) that can have important impacts on behavior. Nonetheless, we are impressed by the fact that ambiguity is a significant aspect of much economic activity and believe that the topic merits greater attention than has been the case to date.

## REFERENCES

- Atkisson, Arthur and Petak, William, *Earthquake Insurance: A Public Policy Analysis*, Report prepared for the Federal Insurance Administration, Federal Emergency Management Agency, Washington, 1981.
- Becker, Selwyn W. and Brownson, Fred O., "What Price Ambiguity? Or the Role of Ambiguity in Decision Making," *Journal of Political Economy*, February 1964, 72, 62-73.
- Bell, David E., "Regret in Decision Making under Uncertainty," *Operations Research*, 1982, 30, 961-81.
- Ehrlich, Isaac and Becker, Gary, "Market Insurance, Self-Insurance and Self-Protection," *Journal of Political Economy*, July-August 1972, 80, 623-48.
- Einhorn, Hillel J. and Hogarth, Robin M., "Ambiguity and Uncertainty in Probabilistic Inference," *Psychological Review*, forthcoming.
- Ellsberg, Daniel, "Risk, Ambiguity, and the Savage Axioms," *Quarterly Journal of Economics*, November 1961, 75, 643-69.
- Hogarth, Robin and Kunreuther, Howard, "Risk Ambiguity and Insurance," Center for Risk and Decision Processes WP 84-10-05, Wharton School, University of Pennsylvania, 1984.
- Kunreuther et al., Howard, *Disaster Insurance Protection: Public Policy Lessons*. New York: Wiley & Sons, 1978.
- Plott, Charles, "Industrial Organization Theory and Experimental Economics," *Journal of Economic Literature*, December 1982, 20, 1485-527.
- Schoemaker, Paul, "The Expected Utility Model: Its Variants, Purposes, Evidence and Limitations," *Journal of Economic Literature*, June 1982, 20, 529-63.
- Smith, Vernon L., "Microeconomic Systems as an Experimental Science," *American Economic Review*, December 1982, 72, 923-55.
- Tversky, Amos and Kahneman, Daniel, "The Framing of Decisions and the Psychology of Choice," *Science*, 1981, 211, 453-58.
- Alliance of American Insurers (American Insurance Association, National Association of Independent Insurers, Mutual Atomic Energy Reinsurance Pool, and American Nuclear Insurers), *Nuclear Power, Safety and Insurance: Issues of the 1980s—The Insurance Industry's Viewpoint*, 1979.

# UNCERTAINTY, BEHAVIOR, AND ECONOMIC THEORY<sup>†</sup>

## Origin of Predictable Behavior: Further Modeling and Applications

By RONALD A. HEINER\*

Economic theory is founded on the assumption that agents act "as if" they are able to maximize according to well-behaved preferences, regardless of how complex their decision problems might be. Consequently, the theory has never investigated the consequences of a genuine gap between an agent's decision-making *competence* and the *difficulty* of a decision problem (called a *C-D* gap). In a recent paper (1983; hereafter called the "Origin" paper), I outlined a general theory for investigating the latter possibility. A major theme was that recurrent pattern in behavior arises because of decision-making uncertainty due to a *C-D* gap; so that uncertainty becomes the basic source of predictable behavior. Here I discuss certain theory topics that were only sketched in the Origin paper, and briefly suggest a few related applications. To pursue these objectives, I begin with a short restatement of the *reliability condition* introduced in the Origin paper.

### I. The Reliability Condition

The following condition determines whether an agent will benefit from allowing itself the flexibility to select an action  $a$  instead of other choosable actions in its behavioral repertoire, denoted  $A$ .

$$(1) \quad \frac{r_a}{w_a} > \frac{l_a}{g_a} \cdot \frac{1 - \pi_a}{\pi_a} = T_a,$$

<sup>†</sup>*Discussants:* Axel Leijonhufvud, University of California-Los Angeles; Mark J. Machina, University of California-San Diego.

\*Member, 1984-85, The Institute for Advanced Study, Princeton, NJ 08540; and Department of Economics, Brigham Young University.

where  $R_a$  = the "right" circumstances for which selecting action  $a$  will either maintain or raise performance compared to that achievable when agents select actions only from  $A - \{a\}$ ,  $W_a$  = the "wrong" circumstances for which selecting action  $a$  will lower performance compared to that achievable when agents select actions only from  $A - \{a\}$ ,  $l_a$  = the "loss" in performance (compared to selecting only from  $A - \{a\}$ ) if action  $a$  is selected under the wrong conditions  $W_a$ ,  $g_a$  = the "gain" in performance (compared to selecting only from  $A - \{a\}$ ) if action  $a$  is selected under the right conditions  $R_a$ , and  $r_a = p(a|R_a)$  and  $w_a = p(a|W_a)$ ;  $\pi_a = p(R_a)$  and  $1 - \pi_a = p(W_a)$ .

The ratio  $r_a/w_a$  measures the "reliability" of selecting action  $a$  under the right instead of the wrong conditions.  $T_a$  measures the minimum required reliability or "tolerance limit" that must be satisfied before allowing flexibility to select action  $a$  will benefit an agent. Note that this condition becomes interesting only when the ratio  $r_a/w_a$  is finite, so that  $r_a/w_a > T_a$  may or may not be satisfied. Note also that if  $r_a/w_a$  is bounded and  $l_a/g_a$  is positive, then  $r_a/w_a > T_a$  will not hold for a sufficiently small but still positive probability of favorable circumstances,  $\pi_a > 0$ .

### II. Conventional Choice Theory is a Special Case

Let us briefly compare reliability theory with the conventional Bayesian decision framework (see my 1985b paper for further discussion). Let  $\gamma(a; x)$  represent the set of consequences resulting from action  $a \in A$  when a message  $x \in X$  is received; where consequences refer to the statistical likelihood of potential outcomes, and actions may



likewise represent randomized strategies over a set of more basic actions. In addition, let  $V(\gamma)$  represent a value function (such as traditional expected utility) which measures the performance derived from different consequences. Conventional decision theory assumes agents behave according to a maximizing decision rule,  $B^*(x)$ , such that for each message  $x$  received  $a = B^*(x)$  if and only if  $V(\gamma(a; x)) \geq V(\gamma(b; x))$  for any other  $b \in A$ .

Now think of this in terms of the probabilities  $r_a$  and  $w_a$ , where  $M_a \subset X$  denotes those messages  $x \in X$  for which the consequences resulting from action  $a$  are maximal as just defined. If  $x \in M_a$ , then  $x$  must also be in  $R_a$  since the consequences from selecting  $a$  are at least as preferred as those resulting from any other action. Conversely, if  $x \notin M_a$ , then some other action will outperform action  $a$ , and thus would be selected by  $B^*(x)$ . Hence, if action  $a$  were selected when  $x \notin M_a$  then performance would drop compared to that achievable with  $B^*(x)$  restricted to  $A - \{a\}$ . The above two implications together imply  $x \in M_a$  if and only if  $x \in R_a$ ; which in turn implies  $M_a = R_a$  and  $X - M_a = W_a$  for any  $a \in A$ . Therefore, since  $B^*(x)$  selects actions if and only if their consequences are maximal we must have  $r_a = p(a = B^*(x) | R_a) = 1$  and  $w_a = p(a = B^*(x) | W_a) = 0$  for all  $a \in A$ .

The above result implies that  $0 < w_a$  and  $r_a < 1$  cannot hold for any action so long as  $B^*(x)$  is used. Yet, as already mentioned, the condition  $r_a/w_a > T_a$  becomes interesting only when the ratio  $r_a/w_a$  is bounded, which is impossible if  $r_a = 1$  and  $w_a = 0$  for all  $a$ . This can also be understood in terms of the probability of Type I and Type II errors (meaning failure to select an action  $a$  when doing so is optimal and still selecting it when doing so is not optimal), denoted  $t_a^1$  and  $t_a^2$ , respectively. Note  $r_a = 1 - t_a^1$  and  $w_a = t_a^2$ , so that  $B^*(x)$  implies  $t_a^1 = 0 = t_a^2$  for all  $a$ . Conventional theory thus represents the limiting case where agents perfectly respond to information.

However, more important than this conclusion is the analytical shift from decision rules like  $B^*(x)$  to the information-response probabilities  $r_a$  and  $w_a$  derived from them (hereafter called "behavior probabilities"). (For further analysis and experiments about

the  $r_a$ ,  $w_a$  (and  $\pi_a$ ) probabilities, as distinct from subjective Bayesian Probabilities, see my 1985a article). Previous inquiry (most notably by Herbert Simon; for a recent summary, see his 1983 study, pp., 12-23) has never been able to identify a modeling alternative of similar generality and rigor to neoclassical choice theory. In contrast, representing optimal decision rules in terms of the behavior probabilities  $r_a$  and  $w_a$  enables one naturally to see that they are at the boundary of a more general domain. Moreover, with these probabilities we can use research from a number of key fields (including major portions of existing risk and statistical theory) to analyze how they change under different conditions (thereby formally modeling behavior away from the boundary defined by traditional maximization).

### III. The General Validity of the Reliability Condition

The reliability condition was only intuitively developed in the Origin paper (for example, no explicit consideration of the size and topology of action, information, and environmental state spaces was given). In fact, the condition is valid under quite general conditions comparable to the most rigorous treatments of conventional choice theory. Let  $\Delta^i(\gamma, \gamma')$  represent a value function that measures the *difference* in performance from consequences  $\gamma$  and  $\gamma'$ , where  $i$  refers to the type of value function used. If  $\Delta^i$  is axiomatized with certain properties (for example, the "independence" axiom) then it will be separable into  $V(\gamma) - V(\gamma')$ , where  $V$  measures the traditional expected utility of different consequences. However, if the more recent "non-expected utility" theories of Machina, Chew, Fishburn, Loomes and Sugden, etc. (see part 5 of Mark Machina, 1983) are used, then  $\Delta^i$  may not be separable into a single  $V$  function.

Let  $\gamma^B$  denote the consequences produced when agents select actions only from a set  $B \subset A$ . It can be shown for any of the  $\Delta^i$  functions mentioned above (see my 1984 paper), that  $\Delta^i(\gamma^A, \gamma^{A-(a)}) > 0$  if and only if  $r_a/w_a > T_a$  is satisfied. Thus, the basic form of the reliability condition holds for any of the  $\Delta^i$  functions (where  $\Delta^i$  affects how the  $l_a$

and  $g_a$  components are measured). Note, however, that we may still have a bounded  $r_a/w_a$  ratio, so that using a  $\Delta^i$  function does not imply agents are able to maximize according to  $\Delta^i$ . That is, *using value functions like  $\Delta^i$  does not imply agents always behave so as to maximize the "posterior" attainable performance conditional on each message received.* Consequently, the reliability condition provides a way of modeling behavior (with the help of value functions like  $\Delta^i$  and behavior probabilities  $r_a, w_a$ ) which no longer requires the maximizing postulate to be used.

#### IV. Formalizing the Relationship between Uncertainty and Predictable Behavior

Next consider how to formally characterize the relationship between uncertainty and predictable behavior. We can precisely measure behavioral predictability with entropy concepts used in statistical mechanics and information-cybernetics' theory. Let  $h_a$  be the probability of selecting an action,  $h_a = p(a)$ . Then the behavioral entropy of potentially selected actions in  $A$  is,  $E^B = -\sum h_a \log h_a$ , for  $a \in A$ . The entropy measure  $E^B$  increases as agents might select more actions, but no single action is frequently chosen (so that higher  $E^B$  means behavior is more difficult to predict; see Claude Shannon and Warren Weaver, 1963, for further analysis of entropy measures).

Next let  $\rho$  denote the set of  $r_a/w_a$  ratios for choosable actions in  $A$  (i.e.,  $\rho = \{r_a/w_a | a \in A\}$ ). The set  $\rho$  describes an agent's reliability in responding to information about when to select particular actions. The limiting case where agents are perfectly reliable is denoted  $\rho^\infty$ , meaning the  $r_a/w_a$  ratios are infinite.  $\rho$  is said to be *bounded* if the reliability ratios of each action do not exceed some finite upper limit  $K$  (i.e.,  $r_a/w_a \leq K$  for all  $a$ ). In the case of  $\rho^\infty$  agents still face "risk" due to imperfect information, but no additional uncertainty is involved. The term "uncertainty" is thus reserved for the case when  $\rho$  is bounded. In addition, let  $\rho^1$  represent the other extreme case where agents are completely unable to distinguish right from wrong conditions for selecting different actions (so that they are equally likely to select actions under either  $R_a$  or  $W_a$ ). Thus,

$\rho^1$  means  $r_a/w_a = 1$  for all  $a$ . The reliability sets  $\rho$  describe a range of uncertainty possibilities, beginning with  $\rho^1$  at one extreme and proceeding through intermediate cases where  $\rho$  is still bounded,<sup>1</sup> finally limiting on  $\rho^\infty$  where only imperfect information remains.

One can prove (see my 1985b paper for more discussion) that  $E^B(\rho)$  approaches zero as  $\rho \rightarrow \rho^1$ ; and conversely that  $E^B(\rho)$  increases beyond any finite bound as  $\rho \rightarrow \rho^\infty$ . The latter result implies that it is the boundedness of  $\rho$  (i.e., the presence of uncertainty) that limits agents' behavioral entropy. We can thus regard uncertainty as the basic source of predictable behavior in the following sense: *recurrent patterns in behavior noticed by an observer (because  $E^B(\rho)$  is bounded) would not have arisen in the first place without uncertainty affecting agents' decisions.*

#### V. Decomposing Reliability into Two Major Stages

Now extend the reliability concepts discussed above by applying them to the overall relationship between the environment and agents' behavior. This can be split into many substages, but only two major ones are discussed here: namely, from the environment to observed information about its true state; and from such information to finally executing agents' behavior.

Previous discussion focused on the second stage, where  $r_a/w_a$  measured the reliability of responding to the right and wrong information,  $R_a$  and  $W_a$ . We can also talk about the right and wrong environmental states for selecting different actions (i.e., those states where selecting action  $a$  would raise or lower performance compared to that achievable when only selecting from  $A - \{a\}$ ). In order to indicate this distinction, let us explicitly superscript  $R_a$  and  $W_a$  with  $S$  when defined in terms of environmental states and similarly with  $X$  when referring to information about the environment. The reliability of agents' behavior in responding to informa-

<sup>1</sup> Bounded reliability sets  $\rho$  may also be understood as a way of formally defining and quantifying Simon's idea of "bounded rationality."

tion is then denoted (for each action  $a$ ) as  $\rho_a^B = r_a^B / w_a^B$ , where  $r_a^B = p(a = B(x) | R_a^X)$  and  $w_a^B = p(a = B(x) | W_a^X)$ . Separate from agents' ability to use information, the reliability of information about when to select a particular action  $a$  is represented by  $\rho_a^X = r_a^X / w_a^X$ , where  $r_a^X = p(R_a^X | R_a^S)$  and  $w_a^X = p(R_a^X | W_a^S)$ . The overall reliability over both stages (given by the ratio  $\rho_a^{XB} = r_a^{XB} / w_a^{XB}$ ) can be shown to have the following general form,

$$(2) \quad \frac{r_a^{XB}}{w_a^{XB}} = \frac{r_a^X(\rho_a^B - 1) + 1}{w_a^X(\rho_a^B - 1) + 1}.$$

Conventional theory represents the limiting case where agents perfectly use information (i.e.,  $\rho_a^B = \infty$  for all  $a$ ); so that  $\rho_a^{XB}$  reduces to  $\rho_a^X = r_a^X / w_a^X$ . This implies agents' overall reliability depends only on the information available to them. On the other hand, the structure of (2) enables one to explicitly analyze the interplay between imperfect information and imperfect response to information. We can thereby retain the essential elements of Bayesian theory, yet enter a previously uncharted domain of inquiry (one that has seemed beyond the limits of formal analysis).<sup>2</sup> In general, note that  $\rho_a^{XB}$  necessarily drops below  $\rho_a^X$  when  $\rho_a^B$  is bounded, converging on 1 as  $\rho_a^B \rightarrow 1$ . The latter must happen no matter how reliable information might be (i.e., no matter how close  $r_a^X$  and  $w_a^X$  approach 1 and 0, respectively).

#### VI. The Complexity of Information: A Basic Tradeoff

Results in cybernetics and information theory imply that in order to fully describe the environment, an information source must have an intrinsic entropy (denoted  $E^X = -\sum h_x \log h_x$ , where  $h_x = p(x)$ ) at least as great as that of the environment. This means that  $E^X$  is a positive function of  $\rho_a^X$ , denoted  $E^X(\rho_a^X)$ . Next combine this with another

principle that agents have greater difficulty in perceiving and interpreting information as its entropy increases. This principle has been extensively tested and confirmed in experimental psychology (see, for example, Wendell Garner 1962, ch. 3), and formally means  $\rho_a^B$  is a negative function of  $E^X$ , denoted  $\rho_a^B(E^X)$ . By combining both of these relationships into  $\rho_a^B(E^X(\rho_a^X))$  and substituting into  $\rho_a^{XB}$  we thus derive the following basic tradeoff: as information more accurately describes the environment its entropy increases and thereby reduces agents' reliability in responding to it. This is implied irrespective of whether there are any costs of obtaining information (as modeled in conventional "search theory").

From this one can prove that agents may be in the immediate presence of a number of *costless* information sources, yet only respond to the less reliable information which thereby has less complexity to interpret. The latter result (as well as the information-complexity tradeoff itself) is one of a number of broadly applicable behavioral principles that can be rigorously derived from reliability formulas like (2) above. In addition, such formulas enable one to harness a large body of analysis already developed in cybernetics and information theory (but which has found little use in conventional decision theory).

#### VII. Brief Applications

I conclude by sketching some applications of the information-complexity tradeoff, first for biology applications and then human applications.

(a) The above implication about ignoring costless information will be more noticeable as agents become less able to reliably interpret increasingly complex messages (so that  $\rho_a^B$  falls more quickly as  $E^X$  increases). A striking class of examples of this are known as *releasing mechanisms* in animal behavior. The following describes a well-known case.

Niko Tinbergen...studied fighting between male stickleback fish.... In the spring the throat and belly of the males become intensely red. It seemed probable, therefore, that the red color was

<sup>2</sup>See the statements by Kenneth Arrow, Robert Lucas, and John Hey at the start of my 1985a article.

an important stimulus. The investigators presented their subjects with a series of models, some quite like actual male sticklebacks except that they lacked the red coloration, and some showing little resemblance to actual sticklebacks except that they were red on the lower surface. The male fish attacked the red-bellied models, despite their unfishlike appearance, much more vigorously than they did the fish-like ones that lacked red. Surely the sticklebacks could see the other characteristics of the models, but they reacted essentially only to the releasing stimuli from the red belly.

[William Keeton, 1976, p. 504]

(b) Formula (2) implies that if  $\rho_a^B$  is bounded then there exists a finite upper bound in behavioral entropy (denoted  $\hat{E}^B$ ) beyond which the reliability condition cannot hold for all actions in an agent's repertoire. If  $\hat{E}^B$  is less than the entropy of environmental states (denoted  $E^S$ ), then it can also be shown that  $\hat{E}^B$  will drop for a sufficient increase in  $E^S$  (i.e., when the environment becomes sufficiently more unpredictable). On the other hand, one can show agents must display more entropy in their behavior in order to maintain any given level of performance in the face of a more unpredictable environment (i.e., higher  $E^S$  requires a higher minimum behavioral entropy, denoted  $\tilde{E}^B$ , in order to maintain a given performance level; see W. Ross Ashby, 1963, ch. 11). The latter can be summarized by a positive relationship  $\tilde{E}^B(E^S)$ . Thus we have another basic tradeoff whereby increasing  $E^S$  will eventually raise  $\tilde{E}^B(E^S)$  above  $\hat{E}^B$ .

Next suppose agents must achieve some minimum performance level in order to maintain competitive survival in the environment. The last conclusion then implies agents will eventually drop out of the environment as its entropy increases. However, we can also use this result in the converse sense by starting with a given  $\hat{E}^B$  and using  $\tilde{E}^B(E^S)$  to determine an upper limit on  $E^S$  consistent with the survival of particular agents (thereby permitting such agents to evolve in the first place). Those environments (or portions of a larger environment) with small enough  $E^S$

for certain agents to evolve can then be defined as the *niches* of such agents. Note how the latter characterization follows directly from reliability principles derived from (2) above.

(c) From this point a number of related features in ecological structure can be pursued. One possibility is to investigate the kinds of competitive interactions between species that would either tightly constrain or quickly increase the environmental entropy mutually faced by the particular species involved (thereby either ensuring or preventing the survival condition  $\tilde{E}^B(E^S) \leq \hat{E}^B$  from being satisfied). Answering this question will enable specific predictions about which kind of niche structures are likely to evolve and which are not. For example, suppose competition between species within a given geographical region quickly increases the environmental entropy they mutually create for themselves as they become sufficiently similar in morphology or other ecologically relevant dimensions. If this is the case, then spatially overlapping niche structures between ecologically similar species will be very unlikely to evolve. This is the dominant pattern evidenced in nonhuman biology.

Now I briefly turn to human agents competing in exchange environments.

(d) Similar to the above examples one can ask what sorts of exchange environments would quickly raise  $\tilde{E}^B(E^S)$  relative to  $\hat{E}^B$  for the agents involved. This might refer to particular markets, or to interdependencies across a number of markets. Answers to these questions will enable predictions about the kinds of intra- and intermarket structures likely to evolve (for example, about where "flex" vs. "fix"-price markets will tend to be located in a large industrial economy with numerous cross connections between producer, consumer, and financial markets). Such predictions will in part depend on how different institutional features affect the complexity of price and quantity signals likely to be transmitted within and between specific markets.

(e) Another possibility involves the complexity of messages transmitted between agents competing in game theory contexts. Conventional models have tended to char-

acterize increasingly sophisticated "optimal" strategies. However, if agents' reliability  $\rho_a^B$  in responding to information decreases with the uncertainty of potentially transmitted messages,  $E^X$ , then noticeably different solution concepts may result. For example, rigid strategies that allow response only to relatively simple messages (such as tit for tat; see Robert Axelrod, 1984) may come to dominate over more sophisticated strategies (that may require agents to respond to messages which they cannot reliably interpret).

(f) Finally, consider briefly the effects of inflation. Again using formula (2) one can show there is a maximum entropy of information (denoted  $\hat{E}^X$ ) beyond which trying to use more complex information will reduce agents' overall reliability  $\rho_a^{XB}$  (thereby reducing their performance). In addition, if we fix  $E^X$  at any given level, then its reliability  $\rho_a^X$  will drop as the environment's entropy  $E^S$  increases. Assume also  $E^S$  exceeds  $\hat{E}^X$ , so that agents are initially using information with entropy  $\hat{E}^X$ . Then suppose something happens to the environment that increases  $E^S$ . The above two principles imply agents' performance must drop as  $E^S$  increases (i.e., greater entropy of information is necessary to maintain its reliability as  $E^S$  increases, yet trying to use more subtle information with  $E^X > \hat{E}^X$  will only further reduce agents' overall reliability; moreover, those information sources with  $E^X \leq \hat{E}^X$  are all less reliable than before). Agents may be able to recoup part of their losses by responding appropriately to information with smaller  $E^X$ , but they can never fully restore their initial performance.

Now apply the above analysis to the case where inflation rises, but in a manner that also increases the entropy of price and quantity relationships in the economy. If this happens, agents' performance achieved from market participation will unavoidably drop. We thus may have a social cost to rising inflation that is not readily apparent with

conventional macro models (which often view inflation as fully "anticipated," or "neutral," or dealt with via "indexing" and other contractual adjustments, etc.; see Axel Leijonhufvud, 1981, chs. 9 and 10).

## REFERENCES

- Ashby, W. Ross, *An Introduction to Cybernetics*, New York: Wiley & Sons, 1963.
- Axelrod, Robert, *The Evolution of Cooperation*, New York: Basic Books, 1984.
- Garner, Wendell R., *Uncertainty and Structure as Psychological Concepts*, New York: Wiley & Sons, 1962.
- Heiner, Ronald A., "The Origin of Predictable Behavior," *American Economic Review*, September 1983, 73, 560-95.
- , "On Reinterpreting the Foundations of Risk and Utility Theory," Working Paper, The Institute for Advanced Study, August 1984.
- , (1985a) "Uncertainty, Signal Detection Experiments, and Modeling Behavior," in R. Langlois, ed., *The New Institutional Economics*, New York: Cambridge University Press, 1985.
- , (1985b) "Predictable Behavior: Reply," *American Economic Review*, June 1985, forthcoming.
- Keeton, William T., *Biological Science*, 3rd ed., New York: Norton, 1976.
- Leijonhufvud, Axel, *Information and Coordination*, New York: Oxford University Press, 1981.
- Machina, Mark J., "The Economic Theory of Individual Behavior toward Risk: Theory, Evidence and New Directions," Report no. 433, Center for Research on Organizational Efficiency, Stanford University, October 1983.
- Shannon, Claude E. and Weaver, Warren, *The Mathematical Theory of Communication*, Chicago: University of Illinois Press, 1963.
- Simon, Herbert A., *Reason in Human Affairs*, Stanford: Stanford University Press, 1983.

# Individual Rationality, Market Rationality, and Value Estimation

By PETER KNEZ, VERNON L. SMITH, AND ARLINGTON W. WILLIAMS\*

Direct experimental tests of expected utility theory (*EUT*), in which subjects are asked to choose among alternative gambles, or to make judgements as to their willingness to pay (*WTP*), and/or willingness to accept (*WTA*) payment for a gamble, have not been kind to *EUT*. As noted in the survey by Paul Slovic and Sarah Lichtenstein (1983), the results of these interrogations are remarkably consistent in a wide variety of contexts and are robust under examinations designed to determine the effect of monetary incentives, experience, and other factors that might have accounted for the discrepancy between subject responses and the predictions of *EUT*. On the other hand, experimental studies of individual and market behavior based upon *EUT* models of market decision making have yielded results showing high consistency with the predictions of these models (see the references in Smith, 1985). Are individual revealed preferences in some market contexts more likely to be "rational" (consistent with *EUT*) than individual responses to choices among alternative prospects?

Several studies designed to solicit *WTP* and *WTA* responses for a variety of goods have found a wide disparity between the "buying price" and "selling price" measures of individual value (see the study and the citations therein by Jack Knetsch and J. A. Sinden, 1984; hereafter K-S). Values of *WTA* obtained in this way are frequently an order of magnitude greater than values of *WTP* although theoretically *WTP* and *WTA*

"should" differ by no more than a presumed "small" income effect. These results should not be dismissed by economists on the grounds of poor subject motivation because the experiments include some (see those in K-S) that have carefully introduced actual monetary payments and cash compensations and have not relied on hypothetical choices.

These results are explained by K-S (and by most of the authors of such studies) in terms of the Daniel Kahneman and Amos Tversky (1982) "framing" paradigm in which people reveal that they are less willing to spend wealth which they consider to be part of their endowment than wealth not so considered. We would emphasize that such behavior is "irrational" only in the narrow sense of *EUT* as a behavioral hypothesis which may not only be a poor predictor of individual choice, but may not be a satisfactory guide to action. For example, the differential treatment of wealth that has become part of one's endowment may have important survival value which is imprinted in decision behavior. However, K-S (p. 508, fn. 3) report one, perhaps significant, experiment in which they do not get the usual *WTA-WTP* disparity. In this experiment, cash payments and offers for a lunch made to respondents entering an office cafeteria showed a much smaller, statistically insignificant, disparity. In this case, the respondents were about to participate in a familiar market for the commodity being evaluated.

On this same theme an important new study by Don Coursey et al. (1984) uses a repetitive series of Vickrey (second price) auctions to determine market *WTA* and *WTP* prices for entitlements to an unfamiliar item, which they compare with hypothetical measures of *WTA* and *WTP*. They find that although individual bids in the first of a sequence of Vickrey auctions show a large *WTA-WTP* disparity, ending bids, after a

\*Graduate Student in Finance, University of Pennsylvania, Philadelphia, PA 19104; Professor of Economics, University of Arizona, Tucson, AZ 85721; Assistant Professor of Economics, Indiana University, Bloomington, IN 47401. Research support from the National Science Foundation is gratefully acknowledged.

series of auctions, yield no such disparity. Their interpretation is that market-like learning experience yields results that are not inconsistent with the "rational" economic model. Since the motivation for such models is to articulate a theory of market behavior, we would argue that the Coursey et al. work raises fundamental questions as to the appropriate context, in which tests of the observable implications of economic theory are to be carried out. In this emphasis we are not denying the reality of what Kahneman and Tversky have called "framing effects." Such effects are identifiable both in the hypothetical and the reward motivated contexts in which they have been measured. What we are questioning is the uncritical *interpretation* of the implications of this research for the theory of markets (Kenneth Arrow, 1982). According to this interpretation, the results of direct studies showing that individual participants are subject to "framing" effects implies that markets are inefficient. We think this interpretation is incorrect (see Smith) for two reasons.

1) It confuses individual "rationality," in the sense of *EUT*, with market rationality, in the sense of allocative efficiency. Market efficiency is a proposition about allocations given demand behavior, while *EUT* leads to propositions about "rational" demand behavior. Thus *EUT* could have very poor predictive power, but given this *EUT*-inconsistent demand environment, markets might be highly efficient. Also, an essential feature of markets is that price is determined by the marginal traders, and the presence of non-marginal traders who are *EUT* irrational does not imply that market prices are *EUT* irrational.

2) Direct studies of framing have concentrated on inconsistencies in individual responses to questionnaires, or in one-shot buying and selling decisions in which individuals do not have the opportunity to participate in an ongoing market. If individuals modify their opinions and their decisions in the light of this experience, these effects will not be reflected in the instruments that have been used in framing studies. We would argue that although such studies have relevance to measuring people's

preference attitudes, they provide no basis for extrapolation to behavior in markets. Most (but not all) experimental markets show some learning effects over time with equilibrium behavior quite different from start-up behavior.

In this paper we report the results of two series of experiments designed to address two research questions: 1) if we ask subjects *WTP* and *WTA* questions concerning a security that yields a random dividend from a known probability distribution, and that will be traded in a market, do these responses, even if inconsistent with *EUT*, provide "good" predictors of the mean market price generated by trading among these same subjects? By a good predictor, we mean that it does at least as well in predicting the mean market price as does the rational expectations equilibrium (*REE*) theory of such a market. 2) If trading is repeated over several consecutive market periods, and we apply the *WTA-WTP* questionnaire instrument between each of these trading periods, do we observe any trend in effect on subject responses in this context?

### I. Experimental Design and Results, Series I

The first group of experiments on which we report were not designed to study *WTP* and *WTA*. These experiments were designed to study the efficient markets hypothesis (or *REE*) in the context of asset trading under double auction contracting rules, where an asset's value is derived from a random dividend distribution (for a more complete discussion, see Smith et al., 1985). In these experiments, nine (or twelve) subjects participate in a sequence of fifteen consecutive trading periods for an asset whose dividend probability distribution is common knowledge. At the end of each trading period, all shares receive the dividend realized for that period. In period 1, each share represents a prospect that will yield fifteen realizations from the given distribution; in period 2 a share can claim fourteen remaining realizations; and so on until the final period when only one realization remains. Each subject begins period 1 with an endowment of cash and securities. The expected (dividend) value

of a share in period 1 is \$3.60 (15 periods times the expected dividend per period, \$0.24), or, in some designs, \$2.40. This is the *REE* price based on expected value as the "intrinsic worth" of a share (the *REE* price adjusted for risk-averse or risk-preferring behavior may be below or above this expected dividend value). Each subject ends the 15-period experiment with his or her initial cash endowment plus the sum of all dividend realizations plus (minus) any net capital gains (losses) from trading shares during the fifteen periods.

After completing an initial series of twelve experiments of this type, we applied a *WTA-WTP* instrument in our subsequent fourteen experiments (Series I). After each subject had completed the asset market instructions and been informed of the dividend structure and of his or her initial endowment, but before the opening of the first trading period, the following two questions were administered: 1) Given your endowment of \$ \_\_\_\_\_ cash (i.e., working capital) and \_\_\_\_\_ asset units, what would be the minimum price you would be willing to accept in order to sell one unit of your inventory in the trading period about to begin? \_\_\_\_\_. 2) Given your endowment of \$ \_\_\_\_\_ cash (i.e., working capital) and \_\_\_\_\_ asset units, what would be the maximum price that you would be willing to pay in order to buy one unit of this asset in the trading period about to begin? \_\_\_\_\_.

The results of our fourteen Series I experiments are shown in Table 1 which lists the buyer and seller surpluses and the predicted market price ( $P_w$ ) based on the *WTA-WTP* interrogations, as well as the actual mean transaction price ( $\bar{P}$ ) for all trades in period 1. Price  $P_w$  equates supply and demand, where supply is the *WTA* responses ordered from lowest to highest, and demand is the *WTP* responses ordered from highest to lowest. Our results are summarized as follows.

1) We first asked if the mean price deviations in period 1,  $p_r = \bar{P} - P_r$  (measured in deviations from the normalizing *REE* price,  $P_r$ ) differ between the prior series of twelve experiments that did not apply the *WTA-WTP* interrogation, and the fourteen Series I experiments that did. Since these sample

TABLE 1—SERIES I RESULTS

Exp.	Surplus Seller : Buyer	<i>WTA, WTP</i> Price	Mean Price
116	2.9:11.1	1.47	1.53
117 <sup>a</sup>	2.1: 8.8	1.00	1.74
118	4.4: 4.0	1.90	2.10
119 <sup>b</sup>	0.5: 2.0	3.00	5.44
125 <sup>b</sup>	2.4: 2.3	2.00	2.55
128 <sup>b</sup>	3.1: 1.1	1.62	1.85
139 <sup>b</sup>	0.6: 4.1	0.60	0.42
141 <sup>a</sup>	4.3:12.3	1.62	2.32
142 <sup>b</sup>	5.2:11.9	2.32	3.78
143 <sup>b</sup>	8.1: 2.9	3.59	3.42
146 <sup>b</sup>	1.4:10.2	1.00	2.67
148 <sup>b</sup>	6.1: 2.8	2.62	3.55
149 <sup>b</sup>	3.8: 2.7	2.75	3.40
150 <sup>b</sup>	2.5: 0.4	3.50	3.40

Note: Exp. denotes experiment.

<sup>a</sup>Denotes  $P_r = 2.40$ , otherwise  $P_r = 3.60$ .

<sup>b</sup>Denotes experienced subjects.

mean deviations are \$0.46 and \$0.70, respectively, and their difference is not statistically significant ( $t = 0.56$ ), interrogation appears not to affect prices.

2) Buyers' surplus tends to exceed sellers' surplus, which implies that there is more diversity among intramarginal limit buy prices than among corresponding limit sell prices. However, this difference is not significant ( $t(26) = 1.6$ ). Since *WTA* and *WTP* reflect expectations about future price increases and the concomitant capital gains, it would appear that such expectations varied considerably among subject trader groups.

3) Comparing the predictive error of the *WTA-WTP* equilibrium price,  $p_w = \bar{P} - P_w$ , with that of the *REE* price,  $p_r = \bar{P} - P_r$ , we reject the null hypothesis that  $\sigma_w^2 = \sigma_r^2$  in favor of  $\sigma_w^2 < \sigma_r^2$  ( $F = 2.45$ ). Hence the *WTA-WTP* data do much better in predicting the mean trading price than the widely touted "intrinsic value" or *REE* theory of stock prices.

4) In each experiment we have complete data on all of the bids and/or offers submitted by each subject, allowing us to compare actual bids with  $WTP_i$  and actual offers with  $WTA_i$ . Examining these data we find that the lowest offer made was below the stated  $WTA_i$  for 14 percent of the subjects, while the highest bid entered was above the



stated  $WTP_i$  for 46 percent of the subjects. We conjectured that buyers were more willing to abandon their  $WTP$ s because they sensed a possible rising market in period 1 and were going for capital gains. Most of the markets would subsequently confirm this prediction (see Smith et al.).

## II. Experimental Design and Results, Series II

In a Series I experiment, the first market period requires evaluating a complicated compound gamble. Furthermore, since it is the first of a 15-period sequence of trading periods, it may be quite important in yielding expectations information for each agent. Hence a subject with well articulated  $WTA$  and  $WTP$  attitudes based only on dividend information, endowment, and "thinking about it," may alter those attitudes drastically upon observing the first few trades that reveal something about the behavior of others. Consequently, our second series (II) of three experiments each consisted of four or six single trading periods for an asset in which all subject endowments were *reinitialized* at the beginning of each period of trading. This pure replication design controls on any first-period trading effects due to capital gains expectations across periods (but not, of course, within a period). In these experiments a single draw at the end of one period of trading is made from a binary probability dividend distribution ( $p_1, d_1; p_2, d_2$ ) = ( $\frac{1}{2}, \$0.50; \frac{1}{2}, 2.00$ ), with expected value \$1.25. If we define  $E$  = (cash, shares) as the endowment vector, there are three agent classes,  $E_1 = (\$4.50, 1)$ ,  $E_2 = (\$3.25, 2)$ , and  $E_3 = (\$2.00, 3)$  with three subjects in each class comprising a nine-trader market. Hence the expected value of each agent's endowment is \$5.75 in each of the independently initialized trading periods in an experiment. Except for the effects of learning, each trading period is a pure replication of the market for a single binary gamble. In this environment, we administer the above questionnaire.

The results of our three Series II experiments consisting of 16 independent single-period markets, each preceded by a  $WTA$ - $WTP$  interrogation, are shown in Table 2. Our results are summarized as follows.

TABLE 2—SERIES II RESULTS

Exp.: Period	Surplus Seller: Buyer	$WTA, WTP$ Price	Mean Price
129:1	1.9:0.9	1.25	1.66
2	1.3:0.9	1.25	1.39
3	0.9:0.5	1.62	1.60
4	1.2:0.3	1.50	1.41
133:1	2.2:2.9	1.25	1.30
2	2.0:0.2	1.42	1.51
3	0.4:0.3	1.47	1.52
4	0.5:0.2	1.50	1.58
5	1.0:0.3	1.50	1.51
6	0.2:0.0	1.50	1.58
137:1	1.9:3.1	1.12	1.43
2	1.2:0.5	1.50	1.49
3	1.0:0.2	1.40	1.40
4	0.5:0.2	1.40	1.26
5	0.7:0.4	1.30	1.18
6	0.3:0.5	1.20	1.21

Note:  $P_r = 1.25$  in all experiments.

1) The  $WTA$ - $WTP$  schedules for both the Series I and II experiments typically reveal frequent inconsistencies with  $EUT$ . For example, in period 1 of experiment 133, subjects 3 and 7 report  $WTP$ s that equal or exceed the maximum possible payoff (\$2). However, these violations of simple dominance are not repeated. In spite of the  $EUT$  inconsistencies, the period 1 market-clearing price of \$1.25 can hardly be classified as irrational.

2) Comparing Tables 1 and 2, both seller and buyer surpluses tend to be smaller in Series II (period 1) than their counterparts in Series I. This is consistent with the expectation that moving from a compound to a single one-draw gamble will reduce the diversity in both the  $WTA$  and  $WTP$  measures of individual value. In Table 2, sellers' surplus is larger than buyers' surplus in all but three trading periods.

3) In ten of sixteen periods  $P_w$  is a better predictor of  $\bar{P}$  than  $P_r$ , but the repeat trials with the same subject group in each experiment yields a sample that is too small (three independent sets of observations) for a meaningful test.

4) If in each trading period we compare  $WTA_i$  with actual offers and  $WTP_i$  with actual bids for each subject, we find that for 34

percent of the subjects the lowest offer made was below their stated  $WTA_i$ , and for 47 percent the highest bid was above their stated  $WTP_i$ . The abandon with which subjects violate their own  $WTA$ - $WTP$  responses suggests that the latter may serve only as a pretrade bargaining objective from which there is frequent deviation.

5) Standard (risk-averse) theory suggests that for any  $i$ ,  $WTA_i$  will equal or exceed  $WTP_i$ . We counted the number of individuals who violated this inequality in their period-by-period  $WTA$ - $WTP$  responses. The number of violations in successive trading periods of each Series II experiment are as follows: Exp. 129: 0,0,0,0; Exp. 133: 3,2,1,2,3,0; Exp. 137: 5,2,1,0,2,1. By this measure of "rationality," it appears that replication of the interrogation followed by market experience sequence tends to reduce the incidence of  $EUT$  inconsistent responses, with most of the reduction occurring in period 2.

### III. Conclusions

The following conclusions appear to be justified by the above experiments: 1) The  $WTA$ - $WTP$  interrogation itself does not have a significant effect on the subsequent observed mean contract price of an asset. 2) In Series I, based on the  $WTA$ - $WTP$  data, buyer's and seller's surplus do not differ significantly. 3) Subjects often submit actual offers below their stated  $WTAs$ , and/or bids above their stated  $WTPs$  (14 to 47 percent of the responses), suggesting perhaps as L. J. Savage once said, "It is difficult to be honest with one's self about prices generally" (1962, p. 165). 4) Although these  $WTA$ - $WTP$  responses are frequently abandoned by subjects, the equilibrium predicted prices,  $P_w$ , that result are not bad predictors of the mean observed contract prices, in the sense that  $P_w$  is closer to the mean price than the "intrinsic" value  $REE$  price,  $P_r$ . 5) Over time in Series II there is a considerable decrease in subject violation of the risk-averse rationality prediction  $WTA \geq WTP$ .

Extensive studies of the consumption-leisure revealed choice behavior of mice, rats, monkeys, pigeons, and people in repeat

purchase environments (Raymond Battalio et al., 1981, p. 623) yield steady-state results consistent with the Slutsky-Hicks model of maximizing behavior. For the animal studies there is no presumption of cognitive calculating choice, yet this presumption is *implicit* in tests of  $EUT$  based on subject responses to one-shot choices among gambles and/or word problems. We do not deny that  $EUT$  is in trouble standing as a clearly nonfalsified theory of decision making under uncertainty, but we would urge suspension of scientific judgement until this evidence has been further examined in repetitive market-like environments. In any case, there is no justification for the normative and judgemental conclusion that  $EUT$  violations imply that either individuals are incompetent or markets are inefficient. What may be in doubt is  $EUT$  as an attempt to give formal meaning to the concept of rationality.

### REFERENCES

- Arrow, Kenneth J., "Risk Perception in Psychology and Economics," *Economic Inquiry*, January 1982, 20, 1-9.
- Battalio, Raymond, Green, Leonard and Kagel, John, "Income-Leisure Tradeoffs of Animal Workers," *American Economic Review*, September 1981, 71, 621-32.
- Coursey, Don, Hovis, John J. and Schulze, William D., "On the Supposed Disparity Between Willingness to Accept and Willingness to Pay," Working Paper, University of Wyoming, 1984.
- Kahneman, Daniel and Tversky, Amos, "The Psychology of Preferences," *Scientific American*, January 1982, 246, 160-73.
- Knetsch, Jack L. and Sinden, J. A., "Willingness to Pay and Compensation Demanded: Experimental Evidence of an Unexpected Disparity in Measures of Value," *Quarterly Journal of Economics*, August 1984, 99, 507-21.
- Savage, L. J., "Bayesian Statistics," in R. E. Machol and P. Grey, eds., *Recent Developments in Information and Decision Processes*, New York: Macmillan, 1962.
- Slovic, Paul, and Lichtenstein, Sarah, "Preference Reversals: A Broader Perspective,"

# Knowledge, Uncertainty, and Behavior

By KEITH D. WILDE, ALLEN D. LeBARON, AND L. DWIGHT ISRAELSEN\*

Our paper reports the initial results of a multidisciplinary evaluation of Ronald Heiner's (1983) thesis about the sources of predictable behavior.<sup>1</sup> A selection of ideas and examples that illuminate or reinforce Heiner's argument is presented here. We also record consensus views on some implications of Heiner's behavioral model as we understand it.

## I. Knowledge/Cognition/Rationality

Our approach to Heiner's thesis is shaped by interest in the processes of cognition, other sources of human behavior, and the implications of knowledge growth for man and the biosphere. His work is illustrative of uncomfortable implications and problems created by the explosive growth of knowledge in the modern era. Many of us work inside large organizations where our responsibility is to promote a kind of intelligent, flexible, and adaptive behavior that Heiner says must emerge as a condition of increased organizational complexity.

From that perspective we were attracted to two of Heiner's major points about human behavior: first, that people cannot cope with all the information available; second,

that knowledge creates uncertainty. We affirm these statements, but are uncertain of Heiner's view on how increasingly complex, viable, social structures evolve. One of our initial impressions was that he attributes consciousness to subhuman forms of life, even though his article claimed to be imputing only sensory or perceptual powers, not cognitive or conceptual ones.

There remains nonetheless an impression that some kind of economic or biological (i.e., success or survival) rationality is the outcome of "successful" behavior. If Heiner is not attributing consciousness, he is at least observing development of behaviors that permit survival. Economic reasoning might call such behaviors "rational" from a retrospective viewpoint. As organisms and organizations become more complex, rationality of this kind seems to require more and more nearly conscious effort. Heiner calls optimization a special case occurring when uncertainty (the *C-D* gap) approaches zero. Optimization would be rational and also conscious, we infer, since it implies deliberate decision taking. At lower levels of certainty, rule-governed behavior prevails. The choice of rules and of behavior within them could be rational without full consciousness. This we infer is Heiner's meaning and we do not necessarily disagree (as some of our examples will demonstrate).

We believe an implication of Heiner's work is that rational (i.e., enabling survival or success) behavior requires increasing degrees of conscious effort. We are given to understand that he is aiming at a more general theory of human behavior, one that subsumes optimization, or economic rationality, as special cases. He aims at illumination rather than revolution he says, but we think the implications are revolutionary. If he is not preserving economic man, then he is destroying the Invisible Hand.

From the anthropological perspective, Heiner has situated economic debate squarely into the mainstream of Continental intellec-

\*Wilde: Economist/Strategic Planner, Canada Department of Agriculture, Ottawa, Ontario, K1A 0C5 Canada; LeBaron and Israelson: Resource Economist and Economist, respectively, Utah State University, Logan, UT 84322.

<sup>1</sup>This is the summary of a longer paper by A. H. Esser, Psychiatrist, Editor & Publisher, New York City; R. W. Jackson Physicist/Policy Advisor, Science Council of Canada; S. Miles, Policy Consultant, Toronto; J. Mitchell, Anthropologist/Information Manager, Canada Department of Agriculture; R. A. Schulz, Faculty of Management, University of Calgary; W. H. C. Simmonds, Engineer/Sociologist/Futurist, National Research Council of Canada (retired); G. Sprakman, Organizational Design Analyst, Government of Alberta; J. A. Wojciechowski, Philosopher of Science, University of Ottawa.

tualist thought of the past two decades. He emphasizes the rationality of systems over that of individual optimizers, clearly paralleling the work of the French structural anthropologist Claude Levi-Strauss (and the structuralist movement in general). That is, in Heiner's model, it is the *system* that is rational, not the individual. Individuals cannot optimize; it is beyond their competence. This amounts, as Heiner noted, to dropping the basic rationality assumption from economics. The assumption has been necessary, up to now, because an uncertain subject implied unpredictable behavior. But, if it is uncertainty that produces predictable behavior, then the subrational subject is embraced within a rational (rule-governed) system. This implies a shift in the orientation of economics from the individual maximizer to the consideration of systems. It also reverses the implication of the previous paragraph. If it is the system that is rational, then Heiner is saving the Invisible Hand at the expense of economic man. The crucial link in this paradox is the role of *conscious* effort.

The two previous paragraphs suggest that Heiner may be writing a general theory of society rather than economics (or *more* than economics to be precise). That is, he may be looking at social behavior from a very high mountain top or a so-called God's eye view, and in an evolutionary time scale, rather than in the short-run horizon of an entrepreneur, bureaucrat, politician, or social reformer. If so, what is the importance of his work in the context of contemporary problems? Will it help take the heat off economists for the current state of the world?

Student enrollment in economics and job opportunities for graduates are flat while those for management students have increased dramatically over the past fifteen years. Are other disciplines dealing with reality more effectively than is economics? If economists are going to address broad questions of social theory, as Heiner is doing, other specialists must concentrate on principles and practices of effective and satisfying social action. What will Heiner's model do in terms of suggesting new questions that may help resolve old problems? One thing that an intelligent hierarchy needs is an understand-

ing of which specialists deal with what problems.

Our chief concern beyond the issue of this paper, or of economics generally, is that the *C-D* gap is growing beyond the capacity of any known optimization technique, and that more and more conscious effort is required to deal with the threat to survival that this implies. We believe that economists should be helping to reduce the problems of "info glut" and of decision-making uncertainty. Too often it is simply assumed that someone else will solve the problem. It would be very unfortunate if Heiner's work were to encourage dogmatic slumber and faith in an Invisible Hand at a time when concerted and conscious effort is required to design new institutions that subserve enlightened values and ethics. That conscious effort is necessary is implied, we believe, by Heiner's demonstration in his paper in this session that if an agent or system is to maintain its performance level in the face of an increasingly complex environment, it *must* find reliable ways to use increasingly complex information. This is an example of a problem area that should appeal to economists.

## II. Sources of Predictable Behavior

Behavior is a compound of three main factors: physiological, psychological, and rational. The physiological factors cover basic human needs and are everywhere similar. They are universal, hence predictable. Psychological factors are more complex and may be further subdivided into feelings and desires. Feelings appear to be similar in most persons, races, and cultures. They are near universal, hence predictable.

Desires, however, are more complex. Desires are not stable. They vary over time and are hard to predict. They grow; they can be stimulated. We blow them up; they become obsessions. They are the driving force in society. They also vary from one culture to another, and thus cannot be taken for granted in economic theory. The growth of knowledge and technology provide material for desires to feed upon.

Rational factors, the creations of the thinking brain, are the most change-inducing

and the least predictable. They are also growing, for knowledge is cumulative. The more knowledge we have with which to rationally serve desires, the less predictable is the outcome. The system becomes more and more open for the generation of knowledge is also the generation of ignorance. Each of us is today relatively far more ignorant than was a well-educated person at the dawn of the nineteenth century. And as more and more powerful knowledge becomes available to more and more people, it becomes more and more true that anything can happen. This link between uncertainty and abundance of information is more positive and direct than Heiner has drawn it.

In applying his theory to organizational behavior, Heiner used a biological analog. Biological systems are hierarchal. For a system to be effective or powerful, it must not only have the benefit of specialization of functions, but also many things must be "forgotten" by subunits. Whereas an amoeba must respond to all things connected to its welfare, a multicellular organism admits of specialization and some cells can "forget" about the responsibilities of some of the others. Yet, there must be linkage between them and a kind of intelligence or overview above them. This is the function of the central nervous system (CNS).

Heiner predicts that organizations will become more flexible as they become more complex, just as do other higher forms of life (man having the most sophisticated CNS). Just as an organism must evolve internal rules of behavior if its separate parts are to coexist in harmony and support their common life, so must the organisms that comprise an organization. Social insects, flocks of birds, and herds of mammals are examples, as are tribes among humans. Successful companies and industries also evolve by gradual establishment of heuristic rules of behavior. The organization defines itself by its own experience of what "works" vs. what does not work. This becomes a corporate culture, or knowledge construct. At least part of this might be called rational behavior that is non-conscious.

Ethnic or national cultures evolve in a similar heuristic way. Culture may be de-

scribed as a set of social rules that groups of people live by. These may be discovered by deliberately, methodically breaking them until the unwritten social rule is manifest. Most people are not conscious of the rules (culture) by which they are living. They only know when confronted with aberrant behavior that "we don't do it that way here." We are not conscious of our accent, our culture, or our world view until it is jarred. To the extent that cultures and world views help their bearers to survive, they may be called rational, as discussed above.

From a neurophysiological viewpoint, the process can be explained by "triune brain" theory. Culturally determined behavior (i.e., subconscious) is probably due to human interactions that involve primarily the limbic system of the brain rather than the cerebral cortex. The limbic system governs emotional responses that lead a person to discover for himself behavioral rules of the family, tribe, or other organization.

This subconscious, emotionally centered learning process has probably been a part of social behavior since before the emergence of man. In this century, students of scientific method have abstracted it to a system or principle called conjecture and refutation. The role of emotion in cognition explains why it has taken so long for mankind to become self-conscious of cognitive process: the limbic system attaches emotive value to ideas generated in the neocortex, and the limbic function of socializing overrides neocortical doubt. The consequence of this emotional-cognitive interaction is that group ideologies are hard to break.

This accounts for our frequent blindness to the fact that parts of our personal worldview or professional paradigm are inconsistent with the facts of observation. In the first place, there is the desire to be one with the herd, to stick to group norms. Beyond that, habitual or patterned ideas and behaviors may persist in spite of knowledge to the contrary, because facts and rational arguments may have insufficient emotional impact for our limbic system to "take cognizance" of them.

Another view of predictability comes from management science and is closely related

to economics itself. Predictability entails some stability of behavior patterns. Each of us has a personal stake in having many things remain as they were. There is an economy of decision making in habitual behavior. There is consequently a value to resisting change until the evidence becomes overwhelming. It saves us from the hard work of thinking.

### III. Implications of Unconscious Rationality

In our multidisciplinary interactions, some unease was exposed regarding what seemed a too-easy presumption of rationality by Heiner. Can it emerge automatically when thinking requires such effort? Bertrand Russell is reported to have said that many people would rather die than think. (In fact they do.) Death in preference to thought is a pretty strong aversion. Other trenchant thinkers have observed and marvelled at the phenomenon: Aldous Huxley, Arthur Koestler, Karl Popper, Erich Fromm, José Ortega y Gasset, Edmund Burke, to name a few.

We wonder how far the evolution of intelligent hierarchies can go without conscious thought and deliberate effort? Who is going to put together the rules for the next stage of social evolution? Do economists wish to leave it up to planners, organizational designers, political scientists, lawyers, and legislators? Or are they still promoting the political/economic model invented by eighteenth-century rationalists (liberalism) and adopted in the political constitutions of most Western nations? Economists seem easily to forget that the system of market capitalism is a political contrivance, a deliberate creation, that their own intellectual ancestors built it, and that they are its maintenance men or priestly class. The implicit attitude of most standard economic work is that the system is simply there, and the "scientific" aspect is to "discover" how it works. On the other hand, economists keep up the pressure to encourage everyone to believe in the market metaphor of society—that the pursuit of self-interest is rational for both individuals and for society at large.

What model of society is implied by Heiner's prediction that institutions have to

evolve that allow individuals or groups to know relatively less about each other and the whole? Does he envisage the development of a strong social central nervous system governed by an holistically rational brain at the top? Or is he offering an explanation of the unseen hand that turns atomistic rationality (self-interest) into the good of the whole? Is he suggesting that members of Western cultures have a sixth sense which enables them to understand instinctively without conscious effort, and to reject self-interest actions that would be damaging to the interest of the whole? Will every primitive man touched by the desire for Western goods instantaneously be converted into a similar marvel? Are we talking about economic man circumscribed by an ever-more complex web of regulations and controls, or about economic superman who knows without thinking and automatically selects only those self-serving actions that are also consistent with the welfare of all others? Is system rationality an automatic blessing, or a product of conscious effort?

To say that organizations that survive have adaptation capacity does not warrant a finding or rule that, as organizations become more complex, they also (necessarily and simultaneously) become more flexible. There have been plenty of organizational dinosaurs and there will be many more. An organization may last a long time without being very flexible or adaptable. The corporate culture may be fallible, yet untested by stress. The same is true of ethnic and national cultures. Fallibility shows up when major environmental change imposes severe stress. This has been happening all over the developing world as uncomprehending people demand Western goods but reject the Western culture that makes having them and using them possible.

That institutions must evolve which enable each agent in the society to know less and less about the behavior of other agents, and about the complex interdependencies generated by their interactions, is an implication of Heiner's theory, but the theory does not provide much insight into the mechanism of change in social institutions. We agree with the statement, but fear that it will not happen by assumption. More than ever, and

with increasing speed and intensity, cultural and institutional evolution must be conscious and deliberate. Man lives more and more in a world of his own making and can no longer rely on everything working out for the best.

#### IV. Limitations of the Ultimate Resource

We noted earlier on that the generation of knowledge is also the generation of ignorance. The capacity of our brains to process the information available to them is challenged by the very productivity of the brain itself. *Attention* may be the most limiting resource.

The difficulty of making intelligent decisions, whether personally or in an organization, has changed from ability to get information to one of processing a super abundance. Responsibility for processing and refining information into intelligence may be delegated but advisors become neurotic because they cannot cover the field. This leads to a corollary neurosis in decision makers: they can claim that a decision was rational (optimal) given the information available, but there is a good chance that inconsistent information could have been found, and that some persons may find it after the fact and confront the decision maker with it.

Knowledge and the power it conveys have had the effect of magnifying desires and expectations of people everywhere, yet the complexities and uncertainties manifest in infoglut are limiting our ability to satisfy those desires. The magnification of desires compared to capacity for fulfillment is increasingly dangerous in a world where the disgruntled can exert great destructive leverage against big systems using powerful but small-scale technology. The recent cases of poisoned pharmaceuticals and food products are examples.

Neuroses generated by this growing *C-D* gap have inspired a variety of coping strategies or behavioral rules that include collusion and conglomeration. Government regulation is another, counter, means. This leads to a reaction against complexity and collectivity. The catch is that most developing problems require more knowledge for their solution,

and the knowledge industry requires some kind of collective support, whether by corporate retained earnings or government collected taxes.

Thus, our desires for goods, thrills, power, and freedom are inconsistent with the complexity, discipline, and collectivity that seem necessary to achieve them. The ultimate desire is to have no constraints on the fulfillment of desires. That this may not be possible is hinted in the etymology of economics which suggests the orientation of a householder managing on a limited budget.

During most of the modern era there has been a tension between technological optimists and skeptics. Malthusians and conservationists/environmentalists have pointed to natural limits to human progress. Fred Hirsch called attention to social limits. Cognitive, cultural, and psychological limits are implications of territory that Heiner has opened up. We have serious doubts about the efficacy of what Julian Simon has termed the ultimate resource—unless the mind is turned toward accepting and coping with the limits to desire satisfaction instead of trying always to transcend them.

The ability of persons and societies to act “rationally” in both the economic and neurophysiological senses is decreasing due to the growth of knowledge. The pursuit of short-sighted self-interest (with knowledge power) contributes to danger and ungovernability, to psychosis, police state law and order, and to warfare. Dissatisfied, uncertain and neurotic people are a set up for demagogues and dictators. The rules that we need to adopt as a defense against the *C-D* gap are brotherly love, compassion, and cultural solidarity. (Many students of Public Choice Theory claim, in contrast, that these precise elements can be ignored.) A consciousness of collective values and purposes is required for the intelligent hierarchy. Through their continuing emphasis on the market metaphor as a social ideal, economists have been contributing to social breakdown by persuading people, intentionally or not, to ignore the implications of their personal actions for the common welfare.

Faith in an Invisible Hand is an expression of technological optimism. Economists

are its custodian, but they have never to our knowledge made a convincing case for how and why it will work in the long run. Heiner's work seems to be heading in the direction of showing that it will not.

#### REFERENCE

Heiner, Ronald A., "The Origin of Predictable Behavior," *American Economic Review*, September 1983, 73, 560-95.



## INTERNATIONAL FINANCE<sup>†</sup>

### The Changing Environment of Central Bank Policy

By ALEXANDRE LAMFALUSSY\*

It was more than a year ago that Charles Kindleberger extended to me his flattering invitation to address this joint meeting of the AEA/AFA. At that time we discussed various possible topics, but eventually decided to leave the title as vague as possible, on the grounds that this should allow me to take up the intellectual challenge of speaking on whatever financial crisis might have conveniently cropped up by the time of the meeting. Well, much in line with the fate besetting current economic forecasting, our timing was amiss: no crisis seems to be at hand.

I do not want to take the easy way out by frightening you with possible future crisis scenarios, only to end up by trying to persuade you that (despite the numerous wrongdoings of governments and even the occasionally silly behavior of market participants) the naturally enlightened and effective cooperation between central banks will either avert the crisis or at least contain it. As an alternative to this somewhat uninspiring approach, I propose to offer you a few reflections on some of the more fundamental problems that monetary policymakers are facing today, both domestically and internationally, and for the handling of which they would be delighted to receive from the academic community some operationally usable advice—the stress being on “operationally usable.”

I should like to focus my comments on two points. The first is that the financial

systems of the main Western industrial countries are in the midst of not one but, in some cases, as many as four interconnected evolutionary processes: disinflation; internationalization; innovation; and deregulation. The second point is that unless I am hopelessly behindhand in my readings, economic theory provides us with only limited guidance for managing our monetary affairs in such a complex process of structural adjustment and institutional change; nor can the observation of history give us much help towards understanding a situation which seems to be without precedent.

Let me begin with a few remarks on the management of the disinflation process. Under the impact of concerted anti-inflationary monetary policies initiated in 1979–80, inflation rates have over the last few years been declining more or less rapidly in all industrial countries. With the exception of a few countries they are, however, still at levels which would have been considered alarmingly high during the early 1960's. Moreover, what we know both from survey data and by inference from the level of interest rates suggests that inflationary expectations have been even slower to move downwards. The crux of the matter is that a slow process of disinflation of this kind carries with it, almost by definition, a good deal of uncertainty regarding future inflation rates—otherwise inflation could not be so sticky. This, in turn, implies that a considerable number of market participants are entering into contracts on terms that will inevitably prove costly for them; in other words, we are far from having seen the last of the casualties, either in the field of international lending or domestically, that are the normal corollary of disinflation. At the same time, the very slowness of the

<sup>†</sup> Presented at the AEA/AFA Joint Luncheon.

\*General Manager, Bank for International Settlements, Centralbahnplatz 2, Basel, Switzerland 4002.

process also implies a continued high cost in terms of unused resources and unemployment. For both these reasons, there is the risk of a political reaction against the process of disinflation itself. On the other hand, an anti-inflationary shock treatment might well have been even more painful, with heavy costs being implied in both the short and long run.

These developments raise at least two sets of questions for policymakers. First, is there any practical alternative to slow disinflation? Is "shock treatment" a genuine alternative? Note that history provides us with good examples of quickly successful disinflation only after phases of hyperinflation, not after the sort of long-lasting, creeping inflationary process which has permeated and distorted most of our Western industrial countries over the last fifteen years or more. In the absence of historical precedents, can theory provide any guidance? There have been a few interesting pieces of analysis of the question of shock treatment vs. gradualism, but the academic debate has remained remarkably scant.

Second, on the assumption that the current policy course is the only practicable one, what are its implications for the prudential side of central banking policies? Can manifestations of financial fragility be taken care of by the normal market mechanism, or does their containment require specific lender-of-last-resort intervention by central banks in order to prevent domino effects? Here, too, I would much welcome a wide-ranging theoretical debate on the mechanics of financial adjustment during a slow process of disinflation, as distinct from crisis manifestations at cyclical turning points.

While I could imagine convincing answers to these questions when viewing the process of disinflation within one closed economy, my imagination begins to falter when I look at this process within the framework of the growing internationalization of domestic banking systems. Whatever ratios you care to consider—the share of external claims or liabilities in the total balance sheet, the relative importance of balance-sheet items in foreign currency, the size of income flows derived from international operations—they

all point to a large and increasing international exposure of the domestic banking systems. The story of financial integration is also reflected in the cross-border transmission of interest rate developments. Interest rate parity holds almost instantaneously in the Euro-currency market; but, what is more important, there is growing statistical evidence of strong interconnections, even under floating exchange rates, between interest rate developments in the major domestic markets. Moreover, the fact that a number of countries, and within these countries private firms, are indebted in foreign currencies means that interest rate developments in these currencies can have a totally unexpected impact on the financial ratios of such debtors. In general, financial impulses emanating from the United States are transmitted remarkably quickly to other financial centers, despite fairly generalized floating. Interest rate "de-coupling" has been possible only within certain limits and by certain countries. Similarly, floating has not prevented strong international transmission links via the "real" side of the business cycle either.

In the best academic tradition, much recent research has gone into analyzing the implications of this state of affairs for exchange rate determination and for the international transmission effects of shifts in the policy mix of a large country, in particular of the United States. This research confirms the day-to-day experience of policymakers, namely that in a financially integrated world no country can isolate itself from the others, no matter what its exchange rate regime. To mention just one example, even determined domestic anti-inflationary policies can be thrown off balance by a real effective exchange rate depreciation induced by capital flows. This clearly raises major policy issues to which there are no unequivocal answers, but I certainly have no grounds for accusing academic researchers of any benign neglect of these problems.

I do, however, have the uneasy impression that insufficient academic work has been devoted to analyzing some other implications of international financial integration. One specific problem area concerns the question of whether the growing across-the-border in-

terdependence increases, or, on the contrary, diminishes the fragility of the Western countries' banking systems. More perfect competition would seem to point to greater resilience, that is, to the ability of the system to take care of itself without any lender-of-last-resort intervention. On the other hand, it does not seem evident to me that more active competition in some fields (i.e., internationally), coupled with continued market imperfections in others (i.e., domestically), add up globally to more perfect competition. I shall return to this question shortly, when reflecting on the subject of deregulation. Another much broader area concerns the normative evaluation of the effects of greater financial integration (i.e., of speedier and much larger financial flows) on a world economy in which international direct investment flows remain limited and which at the same time is exposed to increasing trade barriers or to new types of trade distortions (for example, countertrade). This is Bretton Woods turned upside down—a kind of topsy-turviness which, in my physiocratic simplicity, I view with some suspicion.

The third evolutionary process has to do with the accelerating speed of financial innovations, particularly in North America and the United Kingdom, but also in quite a few other countries, though there, perhaps, attended by less publicity. This process is fueled by market participants' desire to hedge against the uncertainty generated by interest and exchange rate volatility (and is thus partly a reflection of inflationary developments), to circumvent regulations or to avoid taxes, to take up opportunities offered by deregulation or new technology, or simply to respond to market pressure. The result is a flow of new instruments and new techniques, and the blurring of dividing lines between institutions as well as between markets.

Central banks operating in such a fluid environment encounter a variety of problems. There is the problem of identifying suitable targets among the monetary aggregates, broad and narrow, and of recognizing circumstances when it seems appropriate to deviate from these targets. At a time when almost all bank liabilities are beginning to carry interest, I fear that the concept of

transactions balances itself may be becoming elusive. Then, second, there are problems related to the narrowly defined monetary control techniques, that is, to the operational methods by which central banks try to hit their targets. Third, central banks would like to know whether and, if so, how the transmission mechanism from these targets to nominal income is affected, for example, by the proliferation of new instruments, the spreading use of floating interest rates or of financial futures.

Fourth, there are the prudential implications of innovation. What should be done, for instance, on a purely technical level, with a number of balance-sheet items listed as contingent liabilities, or with the host of intermediary balance-sheet items classed somewhere between equity and "traditional" liabilities? How should minimum capital ratios be established? Should such ratios be established at all? Are they not going to produce "evasive" innovations? What are the macroeconomic implications of assigning greater control responsibilities to the supervisory authorities? More fundamentally, we should try to assess the systemic effects of the redistribution of risk realized by means of some of these new techniques and instruments. You may argue that when risk-averse market participants shift risks associated with unexpected interest and exchange rate developments onto willing risk takers, everybody is going to be better off. This may well be the case, but increased collective happiness does not necessarily mean greater systemic stability. Or does it?

The difficulties in analyzing these problems and, therefore, in establishing policy-oriented value judgements are aggravated by two aspects of the current trend in innovations. One is that many of them also have an international dimension. Take the example of swapping a fixed interest claim in one currency on a foreign debtor against a variable interest claim in a different currency on a domestic borrower. Note, at the same time, that the legal obligations attached to a swap are so difficult to define, even within one legal system, let alone when several systems are involved, that the word itself cannot be translated unequivocally into the legally very

precise French language. The point is that I am far from sure that all participants in these swaps fully appreciate the commitments they take on. Second, and more importantly, we are confronted here with a continuous *process*, rather than occasional discrete steps followed by a lengthy pause. There is no time for market participants to adjust themselves fully; the process is truly a dynamic one. Take, for instance, the gradual merging of the Euro-bond market with international bank lending, which is progressively eroding the usefulness of traditionally defined international banking statistics and removing the little transparency which we have managed to create in this particular field. What could be the consequences of this vanishing transparency for the decision-making process of market participants or for policymakers?

Let me now say a few words about deregulation—a topic of great interest in this country as well as in others. This, too, is an ongoing process, rather than a quantum jump from a fully regulated to an entirely free financial system. And if we consider the worldwide financial system, it becomes evident that we are condemned to live with a hybrid system even if the legislature of any single country were to accept such a quantum jump—a remote possibility anyway.

What guidance can theory offer to central banks managing their monetary policy or discharging their prudential duties in this environment? Note that the question is not only whether an entirely free financial system is more efficient (whatever that may mean), or more stable, or more easily “controlled” (in the sense of monetary control) than a regulated one. That is an interesting question but one of little immediate practical relevance. What I should like to know is, first, how the *process* of deregulation, with its inevitable lopsidedness and uncertainties as to the next steps, is working out in practice; and, second, how it could be improved. A deregulated world might be better than a fully regulated one, but a lot can happen on our way from the latter to the former.

I apologize for having presented such an indigestible menu of what might look like institutional trivialities, but I think that quite a lot is at stake. I have in mind in particular

the need to preserve the useability of monetary policy as the main macroeconomic policy instrument. The practical or fundamental limitations of fiscal policy have become obvious: with government expenditure absorbing a very high proportion of resources, few Western European countries have any margins of maneuver for stimulatory fiscal policies, while, for reasons that you know only too well, the United States seems to have got stuck in the opposite direction. In such circumstances, impaired useability of monetary policy would surely have to be counted as a social cost to be set against the benefits of innovation and of deregulation in any global cost-benefit analysis.

Those of you who are familiar with ancient writings will by now have discovered my nostalgia for one of Schumpeter's main themes, namely that economic analysis should concern itself with the process of change, with its succession of cumulative or compensating imbalances, rather than with movements around some identifiable state of equilibrium. When I read his writings, more years ago than I care to remember, I hardly understood what he had in mind and dismissed it anyhow because I could not convert it into equations. As a professional participant in the current process of change affecting financial markets, and having to advise central banks on how to operate in such an environment, I am beginning to have an inkling of what he was driving at—although I am less able than ever to put these thoughts into a rigorous theoretical framework. If some of you could, I am sure that practitioners of monetary policy would appreciate it.

In the meantime—“*en attendant*” as we would say more appropriately in French—practitioners will have to continue to practice. They cannot simply resign and take up gardening, much though some of you might wish them to. For my part, in my advisory capacity, I try to prevent them from succumbing to two opposite temptations.

One temptation is to return to complete “*ad hoc*-ry,” that is, to what the French would call “*naviguer à vue*.” This would be a grave mistake. Full discretion cannot counteract uncertainty; in all likelihood it increases it. Rules, be they monetary aggre-

gates or an exchange rate target, are needed to provide some anchor for the wildly fluctuating expectations of market participants; to make monetary policymakers accountable for their action, including their decisions to deviate from predetermined targets; and to give them leverage in their dealings with governments and parliaments.

The other temptation is to retreat into a world of rigid rules. I hope that I have made it abundantly clear why in the present world

environment I do not believe in a monetary policy based on mechanical rules. It is difficult to define such rules; it is sometimes impossible to apply them; and it would often be irresponsible to stick to them.

The road to follow is somewhere in between: rules applied with a pragmatic sense of discretion. Admittedly, this is more easily said than done, but then monetary policy, like all other policies, remains an art not a science.

AMERICAN ECONOMIC ASSOCIATION

---

PROCEEDINGS  
OF THE  
NINETY-SEVENTH  
ANNUAL  
MEETING

DALLAS, TEXAS  
DECEMBER 28–30, 1984

Minutes of the Annual Meeting  
Dallas, Texas  
December 29, 1984

The Ninety-Seventh Annual Meeting of the American Economic Association was called to order by President Charles Schultze at 5:45 P.M., December 29, 1984 in the Dallas Convention Center. The minutes of the meeting of December 29, 1983 were approved as published in the *American Economic Review, Papers and Proceedings*, May 1984, p. 425.

The Secretary (C. Elton Hinshaw), Treasurer (Rendigs Fels), Managing Editor of the *Journal of Economic Literature* (Moses Abramovitz), Associate Editor of the *American Economic Review* (John Riley), and the Director of *Job Openings for Economists* (Hinshaw) discussed their written reports which were distributed at the meeting. (See their reports published elsewhere in this issue.)

Schultze announced that Orley Ashenfelter had agreed to become managing editor of the

*American Economic Review*, replacing Robert Clower. The transfer of the editorial office will take place in early 1985. He also announced the Executive Committee's decision to initiate steps necessary for the publication of a new journal designed to bring to the general economist the results of frontier research, articles applying theory to policy issues, and information useful to teachers; the new President has been authorized to begin the search for an editor. Schultze then introduced Charles Kindleberger, the 1985 President of the Association, to the audience.

There being no further business, the meeting was adjourned.

Respectfully submitted,  
C. ELTON HINSHAW, *Secretary*



## Minutes of the Executive Committee Meetings

**Minutes of the Meeting of the Executive Committee in Washington, D.C., March 23, 1984.**

The first meeting of the 1984 Executive Committee was called to order at 9:10 A.M. on March 23, 1984 in the Conservatory Room of the Washington Hilton Hotel, Washington, D.C. Members present were Charles L. Schultze (presiding), Gardner Ackley, Rendigs Fels, Victor R. Fuchs, Zvi Griliches, Robert L. Heilbroner, C. Elton Hinshaw, Charles P. Kindleberger, W. Arthur Lewis, Janet L. Norwood, and Joseph E. Stiglitz. Present for parts of the meeting were members of the Nominating Committee: William J. Baumol (chair), Clair Brown, Robert M. Coen, John G. Gurley, Stanley Lebergott, Glenn C. Loury, and Richard H. Thaler. Lawrence R. Klein, chair of the Search Committee for the *American Economic Review* Editor, and Katherine K. Wallman, Executive Director of the Council of Professional Associations on Federal Statistics, were present to make reports.

*Minutes.* The minutes of the meeting of December 27, 1983 were approved as written and circulated prior to the meeting.

*Report of the Secretary* (Hinshaw). The Secretary reported that the 1984 annual meetings will be held in Dallas, Texas on December 28–30, and the next meeting of the Executive Committee will be December 27. The sites for subsequent meetings are New York (1985) and New Orleans (1986). Boston and Chicago are under consideration as sites for the 1987 meetings.

Registrations for the 1983 ASSA meetings in San Francisco totaled 5,735. Forty associations organized 390 scholarly sessions. In 1974 when the meetings were last held in San Francisco, 5,431 people registered and 21 associations organized 196 sessions. The Secretary discussed some of the difficulties the growth in sessions caused for planning, organizing, and coordinating the meetings. The Executive Committee expressed concern that the quantity of papers and sessions was growing faster than their quality. It recommended that regional associations not be allowed to join the ASSA, and that the AEA

President-elect exercise great care in agreeing to joint sessions with other associations. It was understood that it would be desirable to limit or reduce the total number of joint sessions, possibly imposing a limit of 2 joint sessions per association.

The Secretary announced that a directory of members would be published in 1985 as a special issue of the *American Economic Review*. Because of the peculiarity of postal regulations, it is cheaper to publish articles along with directory information than to publish only directory information. The President was instructed to appoint a committee to select articles for the publication.

*Program for 1984* (Kindleberger). He reported that 39 sessions were complete, 29 virtually complete, and enough loose ends that would probably produce about 10 more. Schultze noted that Milton Friedman had written to him lamenting the past decisions of Presidents-elect not to publish comments in the *Papers and Proceedings*. Kindleberger indicated that the next issue would indicate who the discussant of each paper was and that the comments would be available but not published. It was agreed to continue to leave the decision to publish or not to publish comments in the hands of the President-elect. In response to a question about paying the transportation cost of distinguished speakers and various honorees to the annual meetings, it was decided not to pay such expenses.

*Surplus Committee* (Schultze). There appears to be a great interest in publishing a third journal of a nontechnical nature. He reminded the Executive Committee that it had authorized him to spend up to \$25,000 for a feasibility study. He is still searching for someone to do the study. He thought that it was not now possible to have a "sample, dummy copy" of the proposed journal by December, but that substantial progress would have been made by that time on determining whether to proceed or not.

*Council of Professional Associations on Federal Statistics* (Wallman). Wallman, the Executive Director of COPAFS, reviewed the



highlights of her annual report, a copy of which is available from COPAFS, Suite 440, 806 15th Street, N.W., Washington, D.C. 20005. It was decided to become more involved in and aware of COPAFS activities.

*Nominating Committee* (Baumol). The Electoral College, consisting of the Nominating and Executive Committees meeting together, chose Alice M. Rivlin as the nominee for President-elect and Evsey S. Domar and Albert O. Hirschmann as Distinguished Fellows. Baumol reported the following nominees for other offices: for Vice-President (two to be chosen), Elizabeth E. Bailey, Gerard Debreu, Thomas C. Schelling, and Joseph E. Stiglitz; for members of the Executive Committee (two to be chosen), Marcus Alexis, Alan S. Blinder, Daniel McFadden, and Sherwin Rosen.

*Search Committee for AER Editor* (Klein). Klein reported that the Committee had reduced the number of possible candidates to two or three. He expected to have a recommendation prior to the December 1984 meetings of the Executive Committee.

*Other Business*. A. W. Coats, Crawford D. Goodwin, and Laurence S. Moss had written requesting the Association to provide an initial grant to support a systematic investigation of the available archival sources within the United States, especially the unpublished papers and correspondence of leading American economists. It was decided (a) to give general approval of the undertaking, (b) to support the application for a grant from an appropriate body after the completion of an acceptable pilot study, (c) but not to provide funds for the project at this stage.

William D. Gunther, Executive Secretary – Treasurer of Omicron Delta Epsilon, had written asking the Association to share in financing and sponsoring a booklet on careers in economics. Interest was expressed in helping with such a project, but several objections were raised concerning the submitted draft. It was indicated that the Association would participate if a more appropriate document was developed.

Samuel H. Williamson and James Dunlevy, both of Miami University, had written recommending that the Association appoint a committee to establish a standard procedure to be used in generating departmental rank-

ings. It was agreed not to become involved in attempts to determine departmental rankings.

*Report of the Editor of the Journal of Economic Literature* (Moses Abramovitz). Abramovitz wrote that the transition in the Pittsburgh office from Naomi Perlman to Drucilla Ekwurzel was going smoothly. The 1979 edition of the *Index of Economic Articles* will go to press by June and will be published in the fall. The 1980 edition should go to the printer by the end of the year. Abramovitz and Asatoshi Maeshiro, who now has full responsibility for the classification of articles for the *JEL*, are beginning a review of the list of journals indexed by the *JEL*. No such review has been made since the *JEL* began.

*Report of the Treasurer* (Fels). Using the investment income formula adopted in the late 1960's, the Association's audited financial statements show a surplus for 1983 of \$245 thousand. The surplus would have been \$437,440 if changes in the market value of stocks, after adjustment for inflation, had been recognized only but entirely in 1983. In the absence of decisions to undertake new projects or cut dues, the ratio of net worth to expenses is expected to continue to rise. The current ratio is 1.48.

The current dues structure discriminates (unintentionally) against non-academic members. Academics who are Associate or Assistant Professors may pay lower dues than non-academics even though their incomes are more or less the same since dues are based on either of two criteria, rank or income. The Budget and Salary Committee recommended the abolition of the rank criterion. It was VOTED that the Treasurer should draft an amendment to the bylaws abolishing the rank criterion and adjusting the income categories in such a way as to maintain existing revenues. The draft would be circulated to the Executive Committee for its approval and then submitted to a vote of the membership.

There being no further business the meeting adjourned at 4:30 P.M.

*Minutes of the Meeting of the Executive Committee in Dallas, Texas, December 27, 1984.*

*Call to Order*. The second meeting of the 1984 Executive Committee was called to

order at 10:10 A.M. on December 27, 1984 in the Garden Room of the Fairmont Hotel, Dallas, Texas. Members present were Charles L. Schultze (presiding), Moses Abramovitz, Gardner Ackley, Rendigs Fels, Victor R. Fuchs, Zvi Griliches, Robert L. Heilbroner, C. Elton Hinshaw, Charles P. Kindleberger, W. Arthur Lewis, Janet L. Norwood, A. Michael Spence, and Joseph E. Stiglitz. Also present were recently elected members of the 1985 Executive Committee (Elizabeth E. Bailey, Daniel McFadden, and Alice Rivlin) and Counsel Leo Raskind. John Riley, Associate Editor of the *American Economic Review*, sat in for Robert Clower. Present as guests for parts of the meeting were Barbara Bergmann, Donald J. Brown, Lee Hansen, David Richardson, and Bernard Saffran.

In his opening remarks, President Schultze welcomed the new members and thanked those leaving the Committee (Ackley, Griliches, Ann F. Friedlaender, and Heilbroner). He noted that Stiglitz had been reelected to the Committee as a Vice-President. He announced that Orley Ashenfelter had accepted the editorship of the *American Economic Review*; the transfer of the office from Clower would probably take place in March 1985.

*Minutes.* The minutes of the meeting of March 23, 1984 were approved as written and circulated prior to the December meeting.

*Report of the Secretary* (Hinshaw). Hinshaw reported that next year, the AEA's Centennial, the annual meeting would be held in New York during December 27-30; New Orleans would be the site of the 1986 meeting. It was VOTED to accept his recommendation of Chicago as the site for 1987. He announced that a new directory of members is planned as a special December 1985 issue of the *AER*. Since 1985 is the 100th birthday of the Association, a Committee (Moses Abramovitz, Charles Kindleberger, and George Stigler) had commissioned five articles appropriate for the occasion and the directory: "Early American Leaders—The Neoclassical Tradition," James Tobin; "Early American Leaders—The Institutionalist and Critical Tradition," Martin Bronfenbrenner; "The Beginnings of Empirical Economics in America," Carl Christ; "Changes in Methods of Analysis and Research," William

Baumol; and "The Expanding Domain of Economics," Jack Hirshleifer. He also announced that, based on a polling of the Executive Committee during the summer, he had entered into a contract to produce an AEA Centennial T-shirt as part of the 100th birthday celebration. (Samples were available.) It was VOTED to appoint a committee to investigate instituting a journal donation program similar to that of the American Psychological Association.

*Report of the Managing Editor of the American Economic Review* (Riley). Riley, Associate Editor of the *AER*, briefly reviewed Clower's written report (see elsewhere in this issue). He stated that for the first time in several years the number of submissions had not increased significantly; there has been a slight increase in the editorial decision lag over 1983; and about one-third of all submissions are returned without sending them to outside referees.

*Report of the Managing Editor of the Journal of Economic Literature* (Abramovitz). Abramovitz summarized his written report (see elsewhere in this issue). He noted the smoothness of the transition of operations in the Pittsburgh office and lauded the work of Drucilla Ekwurzel, who replaced Naomi Perlman as Associate Editor, and Asatoshi Maeshiro, Professor at the University of Pittsburgh, who serves as an Editorial Consultant with general responsibility for the classification of articles. Among other things, the new team has expanded bibliographic services of the *Journal*, published the 1979 volume of the *Index of Economic Articles* in November, and added abstracts of articles to the material available by computer through DIALOG.

It was recommended that more publicity be given to the availability of the *Indexes* and that the editor consider publishing an article on the use of on-line bibliographic services such as DIALOG.

*Report of the Director of Job Openings for Economists* (Hinshaw). See his written report elsewhere in this issue.

*Report of the Treasurer* (Fels). Fels reported (see his written report elsewhere in this issue) that the projected 1985 budget as circulated needed revising. The *AER* budget contained therein was submitted by Clower

and did not include sufficient provision for the cost of transferring the journal to Ashenfelter, the new editor, nor the cost of an expanded editorial office. Neither did it include the Budget Committee's decision to change the method of calculating investment income. The result of these two revisions changed the originally projected surplus of \$91,000 for 1985 to a small deficit. It was VOTED to approve the budget subject to revision, but to delay the change in the method of calculating investment income until the March meeting. The current method of calculating investment income recognizes equity income to include dividends and capital gains (realized or not) adjusted for inflation and spreads the income over three years; income from debt securities is recognized in the year received. The new method proposed by the Budget Committee would calculate investment income as 4 percent of the market value of the portfolio. The Treasurer was asked to compare the results of applying the two methods over the past fifteen years for the March meeting. He was also asked to request Stein, Roe and Farnham to report on their performance relative to that of the stock market indexes.

*Surplus Committee* (Schultze). Schultze reminded the Executive Committee that he had been authorized to hire a consultant to investigate the feasibility and costs of publishing a third journal. Bernard Saffran had conducted the study and reported his findings, copies of which had been circulated. The Surplus Committee had concluded, based on the Saffran report and its own deliberations, that a new journal directed primarily to publishing (1) frontier research in nontechnical language, (2) applications of theory to policy issues, (3) symposia, and (4) bibliographic material primarily useful to teachers was desirable and feasible. The two most attractive options were an entirely new journal and the division of the *Journal of Economic Literature* into two separate journals—one containing the article and book review sections of the current *JEL* plus the new material and the other containing the bibliographic sections. The Surplus Committee recommended an entirely new journal. Starting such a journal was considered a high risk but worthwhile venture. It was estimated that an

investment outlay of up to \$500,000 might be necessary to launch successfully the journal; this includes substantial outlays on editorial services which are not ordinarily provided by professional journals, sending several issues of the journal free to all members to help establish it in the market place, and up to a year of expenses before the production of the first issue. If successful, it was expected that the new journal could add up to \$200,000 net cost each year to the Association's expenditures. It was anticipated that members would be allowed to pick any two of the three journals (*AER*, *JEL*, and new journal) for their dues and could subscribe to the third for an additional sum. If the experiment was considered a failure, the new journal would be dropped.

Discussion of the recommendation covered the waterfront. One member expressed a feeling of uneasiness about the demand for the proposed journal; another questioned the adequacy of supply of appropriate articles. At least one member preferred to split the *JEL* into two journals; another suggested the possibility of assuming the editorial responsibility of an existing journal, such as the *Journal of Economic Education*, that attempts to address the needs of teachers. Several combinations of format, contents, and size were suggested. No one voiced opposition to the general idea of a third journal; some were enthusiastic supporters.

It was VOTED to publish a new, third journal and authorize the incoming President to seek an editor (or editors) enthusiastic about the concept of a journal directed to the perceived needs of the general economist. It was understood that an initial investment of \$500,000 might be necessary; annual net cost to the Association could be as much as \$200,000; it might take up to a year after an editor is found to produce the first issue; the first several (2 or 3) issues would be sent free to all members; a means of "marketing" the journal would be studied; and the new editor would report to the Executive Committee just as do the current ones.

*Committee on Economic Education* (Hansen). Hansen, who chairs the Committee, reported that there is a growing interest in the teaching of economics at the high school level; thirty states now require precollege

instruction in economics. He sought advice on whether formally to support minimal standards of training, such as at least one college-level course, for teachers. Views expressed on the value of supporting such a standard varied widely—from enthusiastic support to the belief that even minimal standards such as the one given as an example might do more harm than good. No specific advice was given.

*Committee on the Status of Women in the Economics Profession* (Bergmann). In addition to circulating a written report (see elsewhere in this issue), Bergmann stated that the major mission of the Committee was to promote women's participation in all the activities that economists are represented in and to promote getting more women represented in the economics profession. The Committee seeks ways to make the profession more hospitable to women. It was VOTED to appropriate \$15,000 for the activities of the Committee for 1985 and to allow the carryover of some \$13,000 from previous appropriations. It was understood that no "large" expenditure from the carryover would be made without prior approval of the Executive Committee.

*Committee on the Status of Minorities in the Economics Profession* (Brown). Brown reported that the University of Wisconsin hosted its second summer program for minorities in 1984. The Committee has begun the search for a host university for the three years after 1985. Wisconsin has offered to have the program for a fourth year (1986) if necessary. David Richardson briefly reviewed the 1984 program. There were 140 applicants, almost triple the number for the previous year; the 32 applicants accepted were better and more motivated than past classes; 20 of them had the potential to do quality graduate work; 13 or so had the motivation to do so. The 1984 program had been the most successful one to date.

Brown reviewed the status of the Rockefeller grant. In 1980, Marcus Alexis (previous

chair of the Committee) had received a \$350,000 grant to support the summer program, historically black colleges, and graduate fellowships for minorities at a five-school consortium. In 1983, the Federal Reserve System had agreed to support minority graduate students during their third and fourth years given that the Rockefeller grant supported them for the first two years. In part because the Sloan Foundation renewed its grant to support the summer program and because of the initial constraints placed on the granting of fellowships, very little of the Rockefeller monies had been spent by the scheduled expiration date of the grant. Rockefeller was willing to renew the grant provided the five-school consortium was disbanded, any school willing to forego tuition would be eligible for fellowships, and the Federal Reserve System would agree to support students for the third and fourth years. In summary, Rockefeller agreed to support, beginning January 1985, for three years a national fellowship program for minorities. It was VOTED to give approval to Brown to accept the grant on behalf of the AEA provided a satisfactory plan (in detail) for fellowship selection was presented to the Executive Committee at its March 1985 meeting.

*1985 Program* (Rivlin). Rivlin, President-elect and Program Chair for the 1985 meetings, stated that she was planning a special program to celebrate the AEA's Centennial. It would emphasize policy and the practicality of economics. She expected it to have more of a prospective view ("Where do we go from here?") than a retrospective one ("Where have we been?").

There being no further business, the meeting was adjourned.

Respectfully submitted,  
C. ELTON HINSHAW, *Secretary*

## Report of the Secretary for 1984

*Annual Meetings.* In 1985, the Association's centennial, the annual meeting will be held in New York on December 28–30. A special program is planned. The schedule for subsequent meetings is New Orleans in 1986, and Chicago in 1987. Each of these meetings is scheduled for December 28–30, and each will have a Placement Service, which will open for business one day earlier (December 27) than the meetings.

*Elections.* In accordance with the bylaws on election procedures, I hereby certify the results of the recent balloting and report the actions of the Nominating Committee and the Electoral College.

The Nominating Committee, consisting of William Baumol, Chair, Clair Brown, Robert M. Coen, John G. Gurley, Stanley Lebergott, Glenn C. Loury, and Richard Thaler submitted the nominations for Vice-Presidents and members of the Executive Committee. The Electoral College, consisting of the Nominating Committee and Executive Committee meeting together, selected the nominee for President-elect. No petitions were received nominating additional candidates.

*President-Elect*  
Alice Rivlin

<i>Vice President</i>	<i>Executive Committee</i>
Elizabeth E. Bailey	Marcus Alexis
Gerard Debreu	Alan S. Blinder
Thomas C. Schelling	Daniel McFadden
Joseph E. Stiglitz	Sherwin Rosen

The Secretary prepared biographical sketches of the candidates and distributed ballots last summer. On the basis of the canvas of ballots, I certify that the following persons have been duly elected to the respective offices:

*President-elect* (for a term of one year)  
Alice Rivlin

*Vice-Presidents* (for a term of one year)  
Elizabeth E. Bailey  
Joseph E. Stiglitz

*Executive Committee* (for a term of three years)

Alan S. Blinder  
Daniel McFadden

In addition, I have the following information:

Number of legal ballots	5,804
Number of invalid envelopes	167
Number of envelopes received after October 1	24
Number of envelopes returned	5,995

In accordance with the action taken by the Executive Committee at its March 23, 1984 meeting, an amendment to Article I, Section 2, of the bylaws was submitted to the members in a mail ballot in conjunction with the balloting for officers. I certify that the amendment was approved by a vote of 3,460 "for," 601 "against." The bylaws as amended now read:

### Article I, Section 2

There shall be six classes of members other than honorary: regular members with annual incomes of \$30,000 or less paying the base fee defined below; regular members with annual incomes above \$30,000 but no more than \$40,000 paying one and one-fifth times the base fee; regular members with annual incomes above \$40,000 paying one and two-fifths times the base fee; family members (persons living at the same address as a regular member, additional memberships without subscription to the publications of the Association) paying one-fifth of the base fee; junior members (available to registered students for three years only) paying one-half the base fee; and life members comprising those who qualified for life membership by making a single payment of the designated amount prior to January 1, 1976, and exempt from annual fees.

Effective January 1, 1976, the base fee is \$25.00 per year. The Executive Committee may increase the base fee in

proportion to the increase occurring after January 1, 1976 in relevant price and wage indexes. It may increase the income brackets for regular members but may not decrease them below the figures specified in this bylaw.

*National Registry.* The National Registry for Economists continues to be operated on a year-round basis by the Illinois State Employment Service. Economists looking for jobs and employers are urged to register. This is a placement service that maintains the anonymity of employers. The Association is indebted to the Registry for assistance and supervision at the employment service provided at the annual meetings. Employers are reminded of the Association's bimonthly publication, *Job Openings for Economists*, and their professional obligation to list their openings.

*Membership.* The total number of members and subscribers is shown in Table 1. The total has fluctuated between 25,000 and 26,500 since 1975, when it reached an all time high of 26,787. Since then the increase in memberships has offset the decline in subscribers.

*Permission to Reprint and Translate.* Official permission to quote from, reprint, or translate and reprint articles from the *American Economic Review* and the *Journal of Economic Literature* totaled 265 in 1984, compared to 176 in 1983. Upon receipt of a request for permission to reprint an article, the publisher or editor making the request is instructed to obtain the author's permission in writing and send a copy to the Secretary as a condition for official permission. The Association suggests that authors charge a fee of \$150, but they may charge some other amount, enter into a royalty arrangement, waive the fee, or refuse permission altogether.

*1985 Directory.* A new directory of members is planned as a special December 1985 issue of the *American Economic Review*. Since 1985 is the 100th birthday of the Association, articles appropriate for the occasion have been commissioned for the issue. Members should expect to receive biographic questionnaires in early spring.

TABLE 1—MEMBERS AND SUBSCRIBERS  
(End of Year)

	1982	1983	1984
Class of Membership			
Annual	16,771	16,728	16,612
Junior	1,895	1,998	1,932
Life	384	383	370
Honorary	32	31	29
Family	397	423	422
Complimentary	607	599	521
Total Members	20,086	20,162	19,886
Subscribers	5,171	5,986	5,846
Total Members and Subscribers	25,257	26,148	25,732

*AEA Staff.* The Secretary is fortunate to have an unusually loyal and efficient staff. Mary Winer (Administrative Director), Kimberly Adair, Norma Ayres, Ersye Burns, Violet Sikes, Dale Wagner, and Jacquelyn Woods handle the day-to-day operations of the Association. Barbara Weaver and Marlene Hall plan and organize the details of the annual meeting. I wish to thank them for their dedicated work and good humor.

*Committees and Representatives.* Listed below are those who served the Association during 1984 as members of committees or representatives. The year in parentheses indicates the final year of the term to which they have been appointed. On behalf of the Association, I wish to thank them all for their services.

*Ad Hoc Committee on Financial Reporting and Changing Prices*

Franco Modigliani, *Chair*  
Charles Christenson  
Robert Kaplan  
Richard W. Leftwich  
G. William Schwert  
John B. Shoven

*Budget Committee*

Rendigs Fels, *Chair*  
Ann F. Friedlaender, (1984)  
A. Michael Spence, (1985)  
Janet L. Norwood, (1986)  
Charles P. Kindleberger, *ex officio*  
Charles L. Schultze, *ex officio*

*Census Advisory Committee*

Richard E. Quandt, *Chair* (1985)  
Morris A. Adelman (1984)  
Rosanne E. Cole (1984)  
Martin H. David (1984)  
Sidney L. Jones (1984)  
Edwin Mansfield (1984)  
Laurence Chimerene (1985)  
Ronald L. Oaxaca (1985)  
Joel Popkin (1985)  
Ann D. Witte (1985)

*Committee on Economic Education*

W. Lee Hansen, *Chair* (1985)  
Michael K. Salemi (1984)  
John J. Siegfried (1984)  
Kalman Goldberg (1985)  
Campbell R. McConnell (1985)  
Daniel H. Saks (1986)  
Marianne A. Ferber (1986)  
Rendigs Fels, *ex officio*

*Economics Institute Policy and Advisory Board*

Edwin S. Mills, *Chair* (1986)  
Bent Hansen (1984)  
Louis Wells (1984)  
Robert E. Evenson (1985)  
W. Lee Hansen (1985)  
John R. Moroney (1985)  
Dwight Perkins (1987)  
G. Edward Schuh (1987)  
Lance E. Davis (1988)  
Stefan H. Robock (1988)

*Finance Committee*

Rendigs Fels, *Chair*  
Sidney Davidson (1984)  
Robert Eisner (1985)  
Robert J. Genetski (1986)

*Committee on Honorary Members*

Richard A. Musgrave, *Chair* (1986)  
Hendrik S. Houthakker (1984)  
George J. Stigler (1984)  
Hal R. Varian (1986)  
Richard E. Caves (1988)  
Franco Modigliani (1988)

*Committee on Honors and Awards*

Daniel McFadden, *Chair* (1985)  
Oliver E. Williamson (1985)  
Robert Eisner (1987)  
William Vickrey (1987)  
Carlos Diaz-Alejandro (1989)  
Richard R. Nelson (1989)

*Nominating Committee*

William J. Baumol, *Chair* (1984)  
Clair Brown (1984)  
Robert M. Coen (1984)  
John G. Gurley (1984)  
Glen C. Loury (1984)  
Richard H. Thaler (1984)  
Stanley Lebergott (1984)

*Committee on Political Discrimination*

Robert J. Lampman, *Chair* (1986)  
Harold J. Barnett (1984)  
Anne P. Carter (1984)  
Lester C. Thurow (1985)  
Herbert Gintis (1986)  
Richard R. Nelson (1986)

*Search Committee for Editor of American Economic Review*

Lawrence R. Klein, *Chair*  
Robert Eisner  
Claudia D. Goldin  
Robert H. Haveman  
Robert L. Heilbroner  
Joseph A. Pechman  
Joseph E. Stiglitz

*Search Committee for Editor of Journal of Economic Literature*

Gardner Ackley, *Chair*  
Stanley W. Black  
Alan S. Blinder  
James Buchanan  
Albert Rees  
Allen C. Kelley  
V. Kerry Smith

*Committee on the Status of Minority Groups  
in the Economic Profession*

Donald J. Brown, *Chair* (1985)  
Bernard E. Anderson (1985)  
Ronald L. Oaxaca (1985)  
Samuel L. Myers, Jr. (1986)  
Rhonda Williams (1986)

*Committee on the Status of Women in the  
Economic Profession*

Barbara R. Bergmann, *Chair* (1985)  
Nancy Ruggles (1984)  
Gail R. Wilensky (1984)  
Joseph A. Pechman (1985)  
Cordelia W. Reimers (1985)  
Aleta A. Styers (1985)  
Lourdes Beneria (1986)  
Bernadette Chachere (1986)  
Mary Fish (1986)  
Sharon B. Megdal (1986)  
Michelle J. White (1986)

Joan G. Haworth  
Charles L. Schultze, *ex officio*

*Surplus Committee*

Charles L. Schultze, *Chair*  
Gardner Ackley  
Elizabeth E. Bailey  
Ann F. Friedlaender  
Juanita Kreps  
A. Michael Spence  
Joseph E. Stiglitz  
Rendigs Fels, *ex officio*  
W. Arthur Lewis, *ex officio*

*Committee on U.S.-Soviet Exchange*

Franklyn D. Holzman, *Chair* (1987)  
Jennifer F. Reinganum (1986)  
Lloyd G. Reynolds (1986)  
Abram Bergson (1987)  
Joseph A. Pechman (1988)  
Richard N. Rosett (1988)

COUNCIL AND OTHER REPRESENTATIVES

*American Association for the Advancement of  
Science Section K on Social, Economic, and  
Political Sciences*

Roger Bolton (1986)

*American Association for the Advancement of  
Slavic Studies*

Judith Thornton (1985)

*American Council of Learned Societies*

C. Elton Hinshaw (1986)

*Review Board of the American Statistical As-  
sociation-Bureau of the Census Fellowships*

Zvi Griliches

*Consortium of Social Science Associations  
(COSSA)*

Henry J. Aaron  
C. Elton Hinshaw

*Council of Professional Associations on Federal  
Statistics (COPAFS)*

Edward F. Denison (1984)

*Eighth Symposium on Statistics and Environ-  
ment-Steering Committee*

Paul Portney (1984)

*Federal Statistics Users Conference*

Paul Wonnacott (1985)

*Internal Revenue Service Conference-Tax  
Administration Research Strategies*

Harvey Galper (1985)

*International Economic Association*

Kenneth J. Arrow (1984)  
C. Elton Hinshaw (1985)

*Policy Board of the Journal of Consumer Re-  
search*

Louis L. Wilde (1985)

*National Bureau of Economic Research*

Carl F. Christ (1984)

*Social Science Research Council and SSRC  
Committee on Programs and Policy*

Hugh Patrick (1984)

*U.S. National Commission for UNESCO*

Walter Salant (1985)



## REPRESENTATIVES OF THE ASSOCIATION ON VARIOUS OCCASIONS—1984

*Inaugurations*

Robert Hunter Chambers III, Western Maryland College

John M. Jordan

Arthur L. Peterson, Lebanon Valley College

Alan S. Caniglia

James A. Hefner, Jackson State University

Charles P. Kindleberger

Robert Lee Green, University of the District of Columbia

Cleveland A. Chandler

Kenneth L. Perrin, West Chester University

John G. Voyatzis

John B. Stephenson, Berea College

Gus T. Ridgel

Centennial Convocation of Middle Georgia College

Forest J. Denman

## ASSA 1984 CONVENTION COMMITTEE

Joseph E. Burns, *Chair*

Leroy O. Laney, *Vice Chair*

Barbara Weaver, *Convention Manager*

Norma J. Ayres

Louis E. Buck

Thomas E. McMillan, Jr.

William E. Gibson

Edward L. McClelland

Sydney Smith Hicks

Thomas B. Fomby

W. Arthur Tribble

Violet O. Sikes

J. Carter Murphy

Martin T. Katzman

Janet Aschenbrenner

Marlene Hall

C. ELTON HINSHAW, *Secretary*

## Report of the Treasurer for the Year Ending December 31, 1984

The finances of the American Economic Association continue to be robust, though not as robust as projected a year ago. The budget for 1984, shown in the fourth column of Table 1, showed a projected surplus of \$288 thousand resulting from an operating deficit of \$72 thousand and an investment gain of \$360 thousand. It now appears that investment gains for 1984 will be considerably smaller than projected. Final audited financial reports for 1984 will be published in the June issue of this *Review*.

The fifth column of Table 1 shows the proposed budget for 1985 considered by the Executive Committee at its meeting on December 27, 1984. It shows an operating loss of \$151 thousand, an investment gain of \$242 thousand, and a surplus of \$91 thousand. A revised budget will be submitted to the Executive Committee for consideration at its meeting on March 22, 1985. The new budget will include additional expenditures for the *American Economic Review* resulting from appointment of co-editors. It will also show considerably smaller investment gains. The combined effect will be to project a deficit instead of a surplus.

The net worth of the Association, estimated at \$2.6 million on September 30, 1984, is well above what is needed for precautionary purposes. Ten years ago the Association survived a time when the net worth was negative, and for years it was below half of annual expenses. Inasmuch as the Association had a deficit in 1970 about half of its expenditures for that year, I have generally felt that a ratio of net worth to the coming year's spending of one-half was the prudent minimum. With expenditures for 1985 projected at less than \$2 million, a net worth of \$1 million would be adequate. This does not mean that the Association has excess funds of \$1.6 million. If it spends such a sum on start-up costs for the proposed new journal, investment income would fall below the operating loss, resulting in substantial deficits, liquidation of investments, and reduction in the net worth. But the funds available for the new project exceed a million dollars.

In 1983 a Surplus Committee was appointed to consider whether the Association should undertake a new journal. The Executive Committee on December 27 decided to do so and to launch a search for an editor. No provision has been made for such spending on the journal in the 1985 budget.

Recently, the Executive Committee has pursued a policy of not increasing dues rates in nominal terms, with the result that, after adjustment for inflation, real dues have fallen. But it recommended, and the membership approved, a change in the dues structure effective January 1, 1985, designed to correct an inequity without changing the revenues derived from this source. Formerly, there were two criteria for determining the applicable dues rate of a member, rank and income. The two criteria had become badly out of line with each other so that some non-academic members had to pay substantially more than academic members with the same income. To correct this inequity, the criterion of rank has been abolished and the salary brackets adjusted to current realities. Since the result otherwise would have been reduced revenues, the base rate for dues was simultaneously increased.

Even if the Executive Committee had not decided on a new journal, continuing existing policies would result in operating losses growing faster than investment gains. The recent surpluses are giving way to deficits. Soon, members' dues and/or library subscriptions rates will have to be increased. Subscription rates have remained the same since the beginning of 1981. Even at that time, they were not high compared to other journals. Inasmuch as demand is highly inelastic, the Association is in a position to raise more revenue whenever it wishes to do so.

The investment gains shown in Table 1 for 1983 and 1984 are calculated on the basis of the investment income formula adopted by the Association in the 1960's. The formula was based on the belief that (1) the Association could legitimately spend real capital gains, (2) the stock market is so volatile that

TABLE 1—AMERICAN ECONOMIC ASSOCIATION BUDGETS, 1984–85  
(Thousands of dollars)

	First Nine Months (Unaudited)		Full Year		
			Actual	Budgeted	
	1983	1984	1983	1984	1985
REVENUES FROM DUES AND ACTIVITIES					
Membership dues	\$ 564	\$ 580	\$ 758	\$ a	\$ 758
Nonmember subscriptions	471	452	631		631
Subtotal	1,035	1,033	1,389	1,425	1,389
Job Openings for Economists Subscriptions	18	18	27	30	27
Advertising	69	72	98	100	100
Sale of Index of Economic Articles	9	7	61	50	100
Sales of copies, republications, and handbooks	21	21	27	28	25
Sale of mailing list	26	21	35	42	35
Annual meeting	34	21	34	25	3
Sundry	42	36	51	50	50
Total Operating Revenue	1,255	1,229	1,723	1,750	1,729
PUBLICATION EXPENSES					
American Economic Review	370	376	480	495	513
Journal of Economic Literature	495	527	638	755	747
Directory	45	49	60	65	68
Job Openings for Economists	34	35	50	60	55
Index of Economic Articles	5	4	33	26	72
Subtotal	948	993	1,261	1,401	1,455
OPERATING AND ADMINISTRATIVE EXPENSES					
General and Administrative	193	190	298	316	320
Committees	31	35	45	60	50
Support of other organizations	38	48	40	45	55
Subtotal	263	274	382	421	425
Total Expenses	1,212	1,267	1,643	1,822	1,880
OPERATING INCOME (LOSS)	43	(38)	80	(72)	(151)
INVESTMENT GAIN (LOSS)	164	199	165	360	242
SURPLUS (DEFICIT)	\$ 207	\$ 162	\$ 245	\$ 288	\$ 91
Ratio, Net Worth To Annual Expenditures			1.3	1.5	1.4

<sup>a</sup> Revenues for dues and subscriptions were not projected separately.

real capital gains from equities should be "recognized" in the income statement over a period of three years, evening out erratic shifts in stock prices, and (3) the surplus or deficit on the income statement when investment gains are so calculated provides a useful guide for decision making.

Although the principles underlying the investment income formula are sound, we are considering phasing it out. It is too complex; a simpler guide to decision making is needed. Moreover, the experience of the last decade and a half has shown that what matters for decision making is the adequacy of net worth

relative to annual expenditures. If the net worth is high, as is presently the case, deficits are legitimate; if the net worth is negative, as it was ten years ago, a balanced budget is not enough. The last line of Table 1 accordingly shows the ratio of net worth to annual expenditures. (In each case the ratio is based on the net worth at the beginning of the year except that, for 1985, the net worth on September 30, 1984, is used.) It should be kept in mind that the ratio is subject to the vagaries of the stock market.

The investment gains for 1983 and 1984 in Table 1 result from using the investment

income formula. Of the \$199 thousand for the first nine months of 1984, \$158 thousand represents one-third of the real capital gains from equities in 1982 and 1983, recognition of which was postponed to 1984. The projected investment gain for 1985 was calculated by applying the historic real rate

of return on equities (6.5) to the portfolio of equities and adjusting for expected inflation of 4.5 percent the estimated yields from other assets.

RENDIGS FELS, *Treasurer*

## Report of the Finance Committee

The Finance Committee of the American Economic Association met at the Chicago Club, Chicago, Illinois, at noon on December 11, 1984. Present were Committee members Rendigs Fels (Chairman), Robert Eisner, and Robert Genetski. Also present were Harvey Hirschhorn, Robert McNeill, and James Weiss of Stein Roe & Farnham, the investment counsel of the Association. Sidney Davidson was absent.

Mr. McNeill presented a written report which included the minutes of selected past meetings charts and tables relating to market performance, data on the portfolio of the Association, and an economic forecast. For the period from December 1977 to November 1984, the total return on the Association's portfolio was +104.1 percent. On the fixed-income part of the portfolio, the return was +113.6 percent (compared to +58.7 percent for the Salomon High-Grade Index); for the equities part, the return was +98.7 percent (compared to +83.0 percent for the Dow Jones Industrial Average).

For the first eleven months of 1984, the return on fixed-income assets was +12.1 percent (compared to +14.2 percent for the Salomon Index), on equities -10.0 percent (compared to -1.2 percent for the Dow Jones Industrial Average and -20.7 percent for NASDAQ Industrials), and on the total portfolio -3.3 percent. The disappointing performance in 1984 was concentrated in four Stein Roe mutual funds that accounted for about one-third of the portfolio (and over half the equities). While individual securities in the portfolio outperformed the market averages, the funds concentrate on growth

stocks, the kind that fall more than others during periods when most stocks decline. While these funds hurt performance in 1984, they contributed to performance in past bull markets and are expected to do so in the future.

Stein Roe felt that the stock market as a whole was undervalued. Evidence for this view included the increase in "real" profits (corporate profits after taxes net of inventory gains and underdepreciation of fixed assets) in the past year from an annual rate of \$100 billion to nearly \$200 billion and the decline of the S&P 400 *P/E* ratio to little more than 6, the lowest on a chart going back to 1958. Stein Roe felt that its mutual funds would rise more than the market as a whole.

The Committee decided to change its directive to Stein Roe. The directive formerly specified that equities constitute 50 to 75 percent of the portfolio. The Committee changed the range to 40 to 65 percent. It believed that high interest rates and the tax exempt status of the Association made a shift to more fixed-income assets desirable. The Committee left standing the proviso that Stein Roe need not sell equities if the upper limit is exceeded because of a rise in stock prices.

In previous years, the report of the Finance Committee included a table listing the stocks and fixed dollar assets in the portfolio managed by Stein Roe as of the end of the year. It is being omitted this time. Members wanting a copy of the table can get one by writing to the Treasurer.

RENDIGS FELS, *Chairman*

# Report of the Managing Editor

## *American Economic Review*

I have nothing special to report concerning the operations of the editorial office during 1984.

### Operations

The recent history of manuscript submissions and papers published is shown in Tables 1 and 2. We received 921 papers in 1984, approximately the same number as in 1983. This is the first time in several years that the level of submissions has not increased significantly.

The disposition of manuscripts received during 1983 and 1984 is shown in Table 3. The acceptance rate for 1984 has not changed appreciably since last year.

Our file of accepted papers as of December 31, 1984, contained 109 manuscripts; 32 of these will appear in March 1985, 32 in June, and the remainder in September.

Additional information about editorial office processing lags is provided in Tables 4 and 5. The average inventory of manuscripts "in process" has leveled off during the year; this is a result not of faster reviewing, but rather of the decline in submissions during the year. There has been a slight deterioration since last year in the distribution of editorial decisions lags (Table 4); that is, processing was completed on only 67 percent of manuscripts submitted as compared to 77 percent in 1983.

The subject matter distribution of papers published in 1983 and 1984 is shown in Table 6. It is my impression that the distribution of published papers reflects fairly accurately the distribution of papers submitted; however, our record keeping procedures do not permit me to confirm or disconfirm this conjecture.

TABLE 1—MANUSCRIPTS SUBMITTED  
AND PUBLISHED, 1965–84

Year	Submitted	Published	Ratio of Published-to-Submitted
1965	420	59	.14
1966	451	62	.14
1967	534	94	.18
1968	637	93	.15
1969	758	121	.16
1970	879	120	.14
1971	813	115	.14
1972	714	143	.20
1973	758	111	.15
1974	723	125	.17
1975	742	112	.15
1976	695	117	.17
1977	690	114	.17
1978	649	108	.17
1979	719	119	.17
1980	641	127	.20
1981	784	115	.15
1982	820	120	.15
1983	932	129	.14
1984	921	138	.15

TABLE 2—SUMMARY OF CONTENTS, 1983 AND 1984

	1983		1984	
	Number	Pages	Number	Pages
Articles	52	747	52	721
Shorter Papers, including Comments and Replies	77	383	86	358
Dissertations		24		22
Announcements and Notes Section		48		71
Index		10		10
Total		1212		1182

TABLE 3—DISPOSITION OF MANUSCRIPTS,  
1983 AND 1984

	1983	1984
Manuscripts Received	932	921
Completed Processing:	732	705
Accepted	86	94
Rejected	646	611
Acceptance Rate	9.2%	10.2%
Currently in Process	200	216

TABLE 6—SUBJECT MATTER DISTRIBUTION OF  
PUBLISHED MANUSCRIPTS, 1983 AND 1984

	Published	
	1983	1984
General Economics and General		
Equilibrium Theory	12	14
Microeconomic Theory	15	13
Macroeconomic Theory	10	7
Welfare Theory and Social Choice	8	10
Economic History, History of		
Thought, Methodology	6	10
Economics Systems	4	2
Economic Growth, Development,		
Planning, Fluctuations	5	2
Economic Statistics and		
Quantitative Methods	5	9
Monetary and Financial		
Theory and Institutions	9	6
Fiscal Policy and Public Finance	8	8
International Economics	6	10
Administration, Business Finance	7	7
Industrial Organization	11	11
Agriculture, Natural Resources	4	8
Manpower, Labor Population	12	13
Welfare Programs, Consumer		
Economics, Urban and		
Regional Economics	7	8
Total	129	138

TABLE 4—DISTRIBUTION OF EDITORIAL DECISION LAGS  
BETWEEN RECEIPT AND REJECTION,  
NOVEMBER 1, 1983—OCTOBER 31, 1984

Weeks to Rejection	Number of Manuscripts	Percent
0-4	300	.49
5-9	113	.18
10-14	92	.15
15-19	59	.10
20-24	22	.04
25-29	7	.01
30+	18	.03
Total	611	100.

TABLE 5—AVERAGE 1984 PUBLICATION LAGS,  
BY JOURNAL ISSUE

1984 Issue	Number of Weeks Lag		
	Receipt to Acceptance	Acceptance to Publication	Receipt to Publication
March	19	37	56
June	18	42	60
September	14	44	58
December	25	46	71

**Expenses: Printing and Mailing**

Table 7 shows the printing and mailing expenses for the four regular issues and for the *Papers and Proceedings*, issue of the *Review* for 1984. As in earlier years, the *Papers and Proceedings* accounted for approximately 25 percent of total printing and mailing expenses. There have been only minor changes in costs during the preceding year, and I expect that situation to continue.

TABLE 7—COPIES PRINTED, SIZE, AND COST OF PRINTING AND MAILING, 1984 AER

	Copies Printed	Pages		Cost		
		Net	Gross	Issue	Reprints	Total
March	28,000	267	312	\$48,280.26	\$1,727.01	\$50,007
May	28,000	473	496	72,672.27	2,465.14	75,137
June	27,500	282	304	49,016.74	1,457.65	50,474
September	27,500	332	344	56,538.30	1,605.22	58,144
December <sup>a</sup>	27,500	301	368	58,327.93	1,500.00	59,828
Annual Misc. <sup>b</sup>						8,960
Total		1,655	1,824	\$284,835.50	\$8,755.02	\$302,550

<sup>a</sup> Estimated.<sup>b</sup> Estimated—Based on costs of preparing mailing list, extra shipping charges, and storage costs of back issues.

Accordingly, expenses for 1985 are projected to be much the same as in 1984.

### Papers and Proceedings

The seventh volume of the *Papers and Proceedings* to be prepared by the editorial staff of the *Review* appeared in May 1984. As in 1983, this task was handled by John Riley, the associate editor, and by Wilma St. John and Theresa De Maria. As in 1983, manuscript was processed directly to page proofs for the 1984 *Proceedings*.

### Board of Editors

The Board of Editors now consists of twelve members, chosen by the managing editor with the approval of the Executive Committee of the Association. As in earlier years, the Board has been helpful in dealing with comments on published articles and in reading papers that are the subject of complaint over the fairness or competence of referees. I am grateful to members of the Board for their assistance and advice, for their generally supportive attitude, and for useful criticism of particular actions of the managing editor or particular aspects of the operations of the *Review*.

Six members of the present Board will complete their terms as of March 31, 1985: George Akerlof, Patricia Danzon, Jack Hirshleifer, Rick Mishkin, Sherwin Rosen,

and Richard Schmalensee. I am most grateful to each of them. I also want to thank the continuing members of the Board for services performed and in prospect: Clive Bull, Michael Darby, Philip Graves, Meir Kohn, Susan Woodward, and Leslie Young.

### Appointment of New Managing Editor

In the spring of 1985 I shall be completing my fifth and final year as managing editor. Orley Ashenfelter has accepted the invitation of the Executive Committee to serve as the new managing editor. Under his leadership, there will be three co-editors who will share in the responsibility of final decisions on manuscripts. Given the lag between acceptance and publication, the first manuscripts to be accepted under the new regime should appear in June or September of 1986.

### Acknowledgements

I wish to thank the associate editor, John Riley, and my office associates, Wilma St. John, Theresa De Maria, Marcus Hennessy, and Beverly Bardi for dedicated performance that is frequently beyond the call of duty. My thanks also to our graduate mathematics consultants Fred Luk and Mario Rui Pascoa. Finally, and most wholeheartedly, I want to express my gratitude to the many referees who have devoted so much time and energy to the advancement of our science.

A. B. Abel  
K. Abraham  
D. Abreu  
I. Adelman  
M. Adelman  
R. Aiyagari  
J. Aizenman  
G. Akerlof  
M. A. Akhtar  
A. Alchian  
B. T. Allen  
J. G. Altonji  
S. Altug  
J. E. Anderson  
E. Applebaum  
C. Archibald

K. Arrow  
C. Ash  
O. Ashenfelter  
A. B. Atkinson  
R. Ayanian  
C. Azariadis  
E. Bailey  
M. N. Bailly  
B. Balassa  
Y. Balcer  
E. Baltensperger  
P. Bardhan  
A. J. Barkume  
D. P. Baron  
N. Barr  
A. Bartel

W. Baumol  
G. S. Becker  
J. R. Behrman  
D. Belsley  
L. Benham  
A. Ben-Ner  
G. Benston  
T. C. Bergstrom  
B. Bernanke  
E. R. Berndt  
T. F. Bewley  
J. Bhagwati  
G. O. Bierwag  
O. J. Blanchard  
R. Blank  
A. S. Blinder

R. Blitz  
L. E. Blume  
T. E. Borcharding  
M. D. Bordo  
M. J. Boskin  
B. Bosworth  
H. Bowen  
S. Bowles  
R. S. Boyer  
D. Bradford  
J. A. Brander  
W. H. Branson  
R. Brecher  
T. F. Bresnahan  
G. Brock  
D. Brookshire



C. C. Brown	P. A. Diamond	P. Ghemawat	H. M. Hochman
J. Brown	W. E. Diewert	R. J. Gilbert	R. J. Hodrick
E. K. Browning	A. K. Dixit	C. Gilroy	W. Holahan
J. Brueckner	D. Dollar	A. Glazer	S. Hollander
M. Bruno	E. Domar	V. P. Goldberg	C. A. Holt
J. Bryant	R. Dornbusch	A. S. Goldberger	I. Horowitz
C. Bull	M. Dotsey	R. S. Goldfarb	B. Horrigan
J. B. Burbidge	A. Drazen	C. Goldin	P. Howitt
R. Burkhauser	D. Easley	F. Gollop	D. Hsieh
G. T. Burtless	R. Eckaus	R. J. Gordon	C. R. Hulten
P. Cagan	A. Edwards	R. H. Gordon	M. D. Hurd
G. Calvo	S. Edwards	G. Gotz	R. P. Inman
J. Y. Campbell	R. G. Ehrenberg	E. Gramlich	Y. M. Ioannides
M. B. Canzoneri	R. Eisner	C. W. J. Granger	L. F. Jackson
D. R. Capozza	B. C. Ellickson	P. Graves	D. Jaffee
D. Card	D. T. Ellwood	J. A. Gray	J. A. James
T. F. Cargill	M. P. Engers	M. L. Greenhut	G. Johnson
J. A. Carlson	R. F. Engle	B. Greenwald	R. Jones
D. Carlton	W. Ethier	Z. Griliches	P. L. Joskow
J. Carmichael	O. Evans	G. M. Grossman	B. Jovanovic
L. Carmichael	R. E. Falvey	H. I. Grossman	J. Judd
J. Carr	R. C. Fair	S. J. Grossman	J. P. Kalt
W. W. Chang	R. Faith	J. D. Gwartney	M. Kamien
R. S. Chirinko	E. F. Fama	F. Hahn	A. Kane
K. Clark	H. S. Farber	R. Hall	J. Kareken
D. L. Cleeton	J. von Farrell	J. Haltiwanger	M. L. Katz
C. T. Clotfelter	J. A. Fay	J. Ham	M. C. Keeley
W. E. Cole	R. Fearn	D. S. Hamermesh	P. Kehoe
W. S. Comanor	R. Feenstra	J. Hamilton	P. Kewen
J. Conlisk	G. S. Fields	R. Hansen	R. W. Kenny
T. F. Cooley	S. Figlewski	A. Harberger	M. Killingsworth
R. Cooper	S. Fischer	J. R. Harris	K. Kimbrough
R. D. Cooter	F. M. Fisher	M. Harris	M. A. King
T. Copeland	R. P. Flood	D. Harrison	M. A. Klein
B. Cornell	R. H. Frank	G. Harrison	M. Kohn
M. Crain	J. Frankel	O. Hart	R. W. Kopcke
A. Cukierman	H. E. Frech III	M. Hashimoto	R. Kormendi
D. C. A. Curtis	J. A. Frenkel	T. Hatta	E. Koskela
B. Dahlby	J. S. Fried	J. A. Hausman	L. Kotlikoff
C. Dahlman	A. Friedlaender	G. Hay	P. Krugman
P. Danzon	B. M. Friedman	F. Hayashi	T. Kuran
M. Darby	D. Friedman	J. G. Head	M. Kurz
M. H. David	M. Friedman	J. J. Heckman	E. A. Kuska
L. Davis	D. Fudenberg	R. Heiner	J. E. Kwoka
L. De Alessi	D. Fullerton	M. Hellwig	F. E. Kydland
H. De Angelo	G. von Furstenberg	E. Helpman	P. Kyle
A. V. Deardorff	E. G. Furubotn	D. F. Hendry	D. Laidler
S. J. De Canio	F. Gahvari	B. Herrick	M. Landsberger
H. Demsetz	J. Gerson	D. D. Hester	K. Lang
J. Demski	M. Gersovitz	A. Heston	L. Lave
A. De Vany	R. Geske	J. R. Hicks	R. Layard
D. Dewey	J. Geweke	J. Hirshleifer	E. Leamer

K. Leffler	J. C. Moore	R. H. Porter	R. Sexton
A. Leijonhufvud	P. Morgan	R. Posner	C. Shapiro
D. R. Leimer	C. Morris	E. C. Prescott	S. Shavell
H. Leonard	J. Muellbauer	D. D. Purvis	W. Shepherd
J. S. Leonard	D. Mueller	R. Quandt	K. Shepslie
S. F. LeRoy	D. J. Mullineaux	J. M. Quigley	S. Shetty
R. C. Levin	G. Mumy	R. Radner	R. J. Shiller
D. Levine	Y. Mundlak	J. Raisian	J. J. Siegel
F. Levy	P. Murrell	G. Ranis	J. Silvestre
W. A. Lewis	P. Musgrave	A. Raviv	C. A. Sims
S. J. Liebowitz	R. Musgrave	J. Reinganum	K. Singleton
C. M. Lindsay	M. Mussa	M. Reinganum	J. Skinner
C. R. Link	R. F. Muth	G. Rest	J. Smith
P. Linneman	B. Nalebuff	P. Richardson	K. V. Smith
S. A. Lippman	H. Neary	J. G. Riley	R. S. Smith
R. B. Litterman	J. P. Neary	M. H. Riordan	E. Smolensky
S. C. Littlechild	T. Negishi	D. J. Roberts	A. Snow
R. E. Lucas, Jr.	P. Nelson	M. Robinson	K. Sokoloff
S. Lundberg	D. A. Nichols	A. J. Robson	R. M. Solow
R. P. McAfee	W. D. Nordhaus	H. Rockoff	H. Somers
M. E. McBride	G. Norman	K. S. Rogoff	H. Sonnenschein
B. T. McCallum	D. C. North	V. V. Roley	C. S. Spatt
J. McCallum	I. Novos	R. Roll	B. Spencer
R. E. McCormick	W. Oakland	A. Rolnick	D. F. Spulbur
M. J. McKelvey	W. E. Oates	H. S. Rosen	T. N. Srinivasan
R. F. McNown	R. Oaxaca	S. Rosen	R. Startz
P. MacAvoy	G. O'Driscoll	M. R. Rosenzweig	G. J. Stigler
L. J. Maccini	W. Oi	J. J. Rotemberg	J. E. Stiglitz
M. Machina	G. S. Oldfield	A. E. Roth	R. Stillman
J. H. Makin	J. A. Ordovery	R. Rothschild	N. L. Stokey
J. M. Malcomson	A. E. Osborne	D. L. Rubinfeld	C. Stuart
D. Malmquist	J. Ostroy	R. Ruffin	D. B. Suits
R. Manning	A. Oswald	A. Rubinstein	L. H. Summers
S. Margolis	M. Paglin	P. A. Rudd	J. Sweeney
J. Markusen	J. C. Panzar	R. P. Rumelt	G. Tabellini
T. Marsh	G. Papanek	R. Sah	J. E. Tanner
S. Martin	D. O. Parsons	S. W. Salant	J. B. Taylor
A. L. Marty	D. Patinkin	G. Saloner	L. G. Telser
H. P. Marvel	J. Peek	P. A. Samuelson	P. Temin
J. L. Medoff	J. M. Perloff	W. F. Samuelson	R. Thaler
A. H. Meltzer	G. Perry	T. J. Sargent	L. Thompson
M. Melvin	M. K. Perry	R. Sato	T. N. Tideman
H. Mendelson	J. Pesando	T. R. Saving	S. Titman
R. T. Michael	R. S. Pindyck	J. L. Scadding	J. Tobin
P. Mieszkowski	M. Piore	D. T. Scheffman	R. D. Tollison
H. Milde	C. Pissarides	T. Schelling	M. Toma
P. Milgrom	M. Plant	F. M. Scherer	R. Topel
J. Mirrlees	R. D. Plotnick	R. Schmalensee	R. M. Townsend
F. S. Mishkin	C. R. Plott	T. P. Schultz	R. Tresch
O. S. Mitchell	S. Polacheck	T. W. Schultz	G. Tullock
R. Moffitt	A. M. Polinsky	C. L. Schultze	S. Turnovsky
D. E. Moggridge	R. D. Porter	J. Seater	L. Tyson

B. D. Udis	R. Weintraub	J. Wilcox	S. Woodward
M. Ureta	A. Weiss	L. L. Wilde	B. Wright
H. Varian	L. Weiss	O. E. Williamson	G. Wright
W. K. Viscusi	B. A. Weisbrod	R. D. Willig	F. C. Wykoff
X. Vives	M. Weitzman	C. A. Wilson	J. Yellen
P. A. Wachtel	F. Welch	F. A. H. Wilson	J. Yinger
M. Waldman	S. Wellisz	J. R. Wolfe	L. Young
T. J. Wales	E. G. West	K. Wolpin	M. A. Zupin
M. Ward	M. Whinston	P. Wonnacutt	
B. Weingast	B. B. White	S. A. Woodbury	

ROBERT W. CLOWER, *Managing Editor*

## Report of the Managing Editor

### *Journal of Economic Literature*

The *Journal of Economic Literature* consists of seven departments. These include the Articles and Book Reviews, which are edited in the Stanford office, and five bibliographic departments, which are edited in the Pittsburgh office. The latter include the basic Subject Index of the articles published in over 270 economics journals, as well as an Author Index and Abstracts of articles from many journals. The Pittsburgh office also edits the annual *Index of Economic Articles* which provides annual cumulations of the subject and author indexes of articles carried in the *Journal* itself (by year of publication in the originating journal), as well as articles published in many collective volumes such as conference proceedings, *festschriften*, and the like. Since 1983, the contents of the annual *Indexes*, brought up to date quarterly, have also been available by computer through the DIALOG Information Retrieval Service. A Note describing the mode of computer access through DIALOG to articles in the *JEL* subject index and the *Index of Economic Articles* appears in each issue of the *Journal* immediately following the Table of Contents.

The Managing Editor edits the Articles department with the help of Associate Editor John Pencavel, and the Book Review department with the help of Associate Editor Alexander Field. The Assistant Editor in the Stanford office is Anne R. Saldich.

During 1984, the Articles department carried twelve major articles as well as Notes and Communications. The Book Review department published 148 reviews.

The *Journal* again refers readers to the statement of its editorial objectives and policies which appeared in an Editor's Note in the June 1981 issue, page 491. In accordance with these policies, the managing editor commissions the *Journal's* expository, survey and review articles and its book reviews. But the *Journal* welcomes proposals for articles.

With the resignation of Naomi Perlman during the past year, supervision of the Pittsburgh office passed to Drucilla Ekwurzel, who had been Assistant Editor for some years and has now been named Associate Editor. Mrs. Ekwurzel works in association with Professor Asatoshi Maeshiro of the University of Pittsburgh, who serves as Editorial Consultant with general responsibility for the classification of articles on which the *Journal's* subject index depends.

I am glad to report that the bibliographical services of the *Journal* have prospered and expanded under the new editorial team. During the past year, the *JEL* has added 17 journals to the approximately 270 whose contents are classified and indexed. The number of abstracts of articles has also increased. The 1979 volume of the *Index of Economic Articles* was published in November. A program of work has now been instituted that will produce two volumes of the *Index* during 1985 and thereafter. This program of accelerated publication will continue through 1987 when the lag in publication of the annual *Index* will have been reduced to the technical minimum, about 18 months. The material available by computer through DIALOG was also enlarged during the past year by the addition of abstracts of articles. A brochure describing the bibliographical search capabilities of DIALOG and the procedures for gaining access to its *Economic Literature Index* (DIALOG file 139) was made available at the registration desk of the 1984 AEA meeting. We also plan to include this brochure in an early mailing to the Association membership.

The great expansion in the number of journals indexed by *JEL* and the considerable number of new journals appearing each year have made it imperative to put our methods of selecting journals for inclusion in the *JEL* indexes on a better basis. As things now stand, it is difficult to add the many new journals to the indexed list. At the same

time, the question arises whether articles in all the journals now indexed are sufficiently valuable to merit inclusion. We are, therefore, in course of making a systematic survey of the journals now indexed. Our list of journals has been classified by field. Panels of economists who are specialists in these fields have been asked to review the journals in their own fields. They are asked to say which of the listed journals, if any, should be dropped, which journals not now listed might be added, and to give the *Journal* their opinions on a number of matters of general policy. Our survey has so far dealt with journals published in English, and the Editorial Board will use the responses to reevaluate the *Journal's* list. Professor Maeshiro took the principal part in preparing the survey and in compiling and analyzing the responses. A similar survey of journals published in languages other than English is now being made.

On behalf of the Association, I should like to express my warm thanks to the 1984 Board of Editors, who helped plan and review the *Journal's* articles, and to the many economists who served as referees both of proposals and drafts. I acknowledge with special thanks the contributions of the following Board members who completed their terms this year: Edwin Burmeister, Edward M. Gramlich, Sherman Robinson, and E. Roy Weintraub.

Ann G. Vollmer and Anita Makler continued on the *Journal* staff at Stanford throughout 1984. Lyndis Rankin continued with larger responsibilities in Pittsburgh. Patricia Andrews and Elizabeth Thornton joined the Pittsburgh staff early in the year. The *Journal* is grateful for their devoted and effective work.

MOSES ABRAMOVITZ, *Managing Editor*

## Report of the Director

### *Job Openings for Economists*

The number of new jobs listed this year increased 15 percent over last year. This interrupted a three-year decline. In 1983, 1,489 new vacancies were advertised; this year 1,718 new jobs were listed. Both academic and nonacademic listings increased although the percentage increase in non-

academic jobs was greater. Table 1 shows total listings (employers), total jobs, new listings, and new jobs by type (academic or nonacademic) for each issue of *JOE* in 1984.

Universities with graduate programs and four-year colleges continue to be the major sources of job listings. Together they con-

TABLE 1—JOB LISTINGS FOR 1984

Issue	Total Listings	Total Jobs	New Listings	New Jobs
<b>Academic</b>				
February	80	147	70	127
April	56	87	53	82
June	24	40	21	35
August	51	107	51	107
October	156	406	137	374
November	137	295	137	295
December	198	451	90	223
Subtotal	702	1,533	559	1,243
<b>Nonacademic</b>				
February	12	54	9	39
April	22	99	17	78
June	15	52	11	36
August	18	46	16	40
October	28	87	25	83
November	27	95	27	95
December	42	174	22	99
Subtotal	164	607	127	470
<b>TOTAL</b>	<b>866</b>	<b>2,140</b>	<b>686</b>	<b>1,713</b>

TABLE 2—NUMBER AND TYPES OF EMPLOYERS LISTING POSITIONS IN *JOE* DURING 1984

Issue	Four-Year Colleges	Universities with Graduate Programs	Federal Government	State/Local Government	Banking or Finance	Business or Industry	Consulting or Research	Other	Total
February	38	42	3	—	5	—	2	2	92
April	22	34	5	2	4	2	7	2	78
June	12	12	2	1	1	1	9	1	39
August	16	35	3	2	3	2	5	3	69
October	47	109	8	—	4	1	13	2	184
November	47	90	9	1	6	1	9	1	164
December	82	116	13	1	8	—	18	2	240
<b>TOTAL</b>	<b>264</b>	<b>438</b>	<b>43</b>	<b>7</b>	<b>31</b>	<b>7</b>	<b>63</b>	<b>13</b>	<b>866</b>

TABLE 3—FIELDS OF SPECIALIZATION CITED: 1984

Fields <sup>a</sup>	February	April	June	August	October	November	December	Totals
General Economic Theory (000)	73	46	19	55	177	140	221	731
Growth and Development (100)	21	17	7	18	44	30	46	183
Econometrics and Statistics (200)	38	36	24	23	68	45	91	325
Monetary and Fiscal (300)	37	30	10	20	97	73	125	392
International Economics (400)	18	20	14	19	45	48	64	228
Business Administration, Finance, Marketing and Accounting (500)	22	13	13	18	48	26	50	190
Industrial Organization (600)	19	25	10	23	57	37	68	239
Agriculture and Natural Resources (700)	10	17	7	13	31	17	40	135
Labor (800)	22	17	8	10	39	35	50	181
Welfare and Urban (900)	14	21	5	12	38	33	51	174
Related Disciplines (A00)	5	2	1	1	8	5	9	31
Administrative Positions (B00)	5	6	3	6	20	8	13	61
TOTAL	284	250	121	218	672	497	828	2,870

<sup>a</sup>Fields of specialization codes are from the *Journal of Economic Literature*.

stitute about 80 percent of total employers. Table 2 shows the number of employers by type for each 1984 issue.

The field of specialization most in demand continues to be general economic theory. Generalists with a strong background in mathematics and statistics appear to be the type of economist that employers are seeking. The applied area of specialization seems to be of secondary importance. Table 3 shows the number of citations by field of specializa-

tion. General economic theory (000) led, followed by monetary and fiscal (300) and econometrics and statistics (200). This pattern has prevailed for the past several years.

Violet Sikes is almost solely responsible for the publication and distribution of *JOE*. I wish to express my great gratitude for the excellent job she continues to do.

C. ELTON HINSHAW, *Director*

## Report of the Census Advisory Committee

This past year the Bureau of the Census implemented important changes in how it utilizes its advisory committees. These changes were adopted in order to reduce duplication of effort by Bureau personnel, and to improve the efficiency and quality of the advice it obtains from its professional advisory committees.

Previously, Bureau personnel met with the various advisory committees on separate dates. In many instances the same briefings were repeated on these separate occasions. Beginning in the spring of 1984, the Bureau started scheduling joint meetings of the Census Advisory Committees of the American Economic Association, the American Marketing Association, the American Statistical Association, and of the Committee on Population Statistics. These committees participate in a joint, plenary session on the morning of the first day of a two-day meeting. For the remainder of the meeting, the

individual advisory committees meet separately or jointly with one or more advisory committees to discuss topics that are more narrowly focused.

In order to bring about greater involvement of the individual committee members and to help focus the discussion, individual committee members are now asked to serve as formal discussants of the briefing papers presented by Bureau personnel. While this had been common practice in other advisory committees, this arrangement had not been followed by the AEA Census Advisory Committee.

Because of budget limitations coupled with a desire to add four new advisory committees for the 1990 Census, the membership on the AEA committee is being reduced through attrition from fifteen members to nine members.

RICHARD E. QUANDT, *Chairman*



## Report of the Representative to the National Bureau of Economic Research

The National Bureau of Economic Research conducts analyses on a large variety of economic issues; publishes books, working papers, and two periodicals; sponsors conferences; and holds workshops and seminars as part of an annual summer institute. Approximately 269 economists at universities across the United States contribute to the NBER's working paper series, and many other economists, here and abroad, attend conferences and the summer institute.

*Programs.* The NBER's research is organized into eight programs (directors in parentheses): Economic Fluctuations (Robert Hall), Financial Markets and Monetary Economics (Benjamin Friedman), International Studies (William Branson), Labor Studies (Richard Freeman), Taxation (David Bradford), Development of the American Economy (Robert Fogel), Health Economics (Victor Fuchs and Michael Grossman), and Productivity and Technical Change (Zvi Griliches). Program meetings are generally held twice during the academic year, and once, for a longer period, during the summer institute. Over 200 people attended summer institute meetings in Cambridge in 1984.

*Projects.* The NBER also sponsors large-scale projects which bring together researchers from several of these programs. John Shoven is overall director of the Bureau's pensions project. David Wise is directing a major study of the effects of pensions on labor market and retirement decisions.

Another major project, centering on the role of Government Budget and the Private Economy, consists of major efforts in the following areas: the impact of taxation on such behavior as charitable contributions (directed by Charles Clotfelter); measuring and analyzing the role of state and local government in the economy (directed by Harvey Rosen); studies of the compensation of public sector employees (directed by David Wise); and the impact of public sector unionization (directed by Richard Freeman);

and an analysis of government debt and deficits and their impact on the private sector (directed by David Bradford and Benjamin Friedman).

A third project, Productivity and Industrial Change in the World Economy, likewise has several major parts. William Branson and J. David Richardson are directing a project on international economic policy. Research on trade relations and trade policy is directed by Robert Baldwin. Richard Marston leads a group studying international macroeconomic coordination. Colin Bradford is directing a study of trade relations with Asian countries. Jacob Frenkel is directing research on exchange rates.

*Conferences.* In 1984, the NBER sponsored conferences (organizers in parentheses) in the United States and abroad on the following topics: Global Implications of the Trade Patterns of East and Southeast Asia (William Branson and Colin Bradford); Economics of Trade Policy (Robert Baldwin, William Branson, and J. David Richardson); Business Cycles (Robert Gordon); Pensions and Retirement in the United States (John Shoven); State and Local Public Finance (Harvey Rosen); International Seminar on Macroeconomics (Georges de Menil and Robert Gordon); International Coordination of Economic Policy (Willem Buiter and Richard Marston); Corporate Capital Structures in the United States (Benjamin Friedman); Public Sector Payrolls (David Wise); and Structural Adjustment and the Real Exchange in Developing Countries (Sebastian Edwards and William Branson); Long-Term Factors in American Economic Growth (Stanley Engerman and Robert Gallman).

*Books.* In 1984, the following NBER books were published by the University of Chicago Press: *R&D, Patents, and Productivity* (Zvi Griliches, ed.); *Economic Transfers in the United States* (Marilyn Moon, ed.); *A Retrospective on the classical Gold Standard, 1821–1931* (Michael D. Bordo and Anna J. Schwartz, eds.); *The Structure and Evolu-*

*tion of Recent U.S. Trade Policy* (Robert E. Baldwin and Anne O. Krueger, eds.); *Exchange Rate Theory and Practice* (John F. Bilson and Richard C. Marston, eds.).

In addition, Harvard University Press published *Economics of Worldwide Stagflation* (Michael Bruno and Jeffrey D. Sachs).

In 1985, the following NBER books will be published: *Social Experimentation* (Jerry A. Hausman and David A. Wise, eds.); *Corporate Capital Structures in the United States* (Benjamin M. Friedman, ed.); *Federal Tax Policy and Charitable Giving* (Charles T. Clotfelter); *A General Equilibrium Model for Tax Policy Evaluation* (Charles L. Ballard, Don Fullerton, John B. Shoven, and John Whalley); *Pensions, Labor, and Individual Choice* (David A. Wise, ed.); *Horizontal Equity, Uncertainty, and Measures of Well-Being* (Martin David and Timothy Smeeding, eds.); *Monitoring Business Cycles in Market-Oriented Countries: Developing and Using International Economic Indicators* (Philip A. Klein and Geoffrey H. Moore).

*Periodicals.* The NBER publishes two peri-

odicals, the *Digest* and the *Reporter*. The *Digest* provides four summaries each month on recent NBER working papers of general interest. The quarterly *Reporter* contains longer summaries of recent program activity, reports of NBER conferences, reviews of recent work by NBER researchers, and abstracts of working papers issued during the previous quarter.

During 1984 Martin Feldstein replaced Eli Shapiro as President of NBER and Geoffrey Carliner replaced David Hartman as Executive Director.

Further information about the NBER is available in NBER publications, from Geoffrey Carliner at NBER, 1050 Massachusetts Avenue, Cambridge, MA 02138, or from the undersigned at Johns Hopkins University, Baltimore, MD 21218.

I am happy to acknowledge Geoffrey Carliner's major contribution to the preparation of this report.

CARL F. CHRIST, *Representative*

## Policy and Advisory Board of the Economics Institute

The year 1984 was a year of recovery and adjustment for the Economics Institute, as it was for the developing countries, from which most of the Institute's students and revenues come.

Total enrollments were up in 1984, following a substantial drop in 1983. The 1984 enrollment was about 1350 students (in 5-week-term student equivalents), about 40 percent each during spring and summer terms, and 20 percent during fall terms.

Staff and administrative stability have benefitted greatly in recent years by the Institute's year-round program. A major innovation in 1983 and 1984 was the operation of pre-Institute training programs in Indonesia. This experiment enabled both the Institute and the sponsors to learn a great deal about operation and administration of such programs, and enabled the Institute to maintain its staff during the recession.

The Institute continues to draw students from throughout the developing countries. The major limitation on its ability to draw large numbers of students from the poorest countries is its need to finance nearly its full costs from tuition and fees, and its corresponding ability to offer only small scholarship assistance. Raising scholarship funds is the Institute's highest priority.

The Board held its annual meeting November 16 and 17, 1984, in Boulder, Colorado. Several Board members also visited the Institute during the summer of 1984. During 1984, Stefan Roback and Lance Davis joined the Board.

WYN F. OWEN, *Director*

EDWIN S. MILLS, *Chairman*

## Report of the Representative to the U.S. National Commission for UNESCO

During 1984, the attention of the U.S. National Commission for UNESCO was dominated by the decision of the U.S. government, announced on December 29, 1983, to withdraw from the organization at the end of 1984. This decision was subject to reconsideration if the practices and administration of UNESCO to which the United States has vigorously objected were significantly changed for the better. The Commission, which consists of private citizens, had earlier expressed general agreement with many or most of the government's criticisms of UNESCO but disagreed with the decision to withdraw, supporting continued U.S. participation largely because it thought the United States in a better position to exert influence toward the needed reforms from inside the organization than from outside it.

As part of the Commission's effort to assist in obtaining the needed reforms and also to observe how far UNESCO was responding to U.S. pressure for those reforms and to observe the attitudes of other member countries toward the U.S. position, six Commissioners attended the meeting of UNESCO's Executive Board in September and October, 1984. The co-chairpersons of this delegation reported to the Commission at its forty-eighth meeting on December 13 that most member countries had become concerned that the United States would withdraw at the end of 1984, irrespective of what changes the organization made during the year. "Most of our allies believed that necessary reforms—many of which they also seek—can best be gained by working from within the organization." By withdrawing, the Commission's observers said, the United States would lose a permanent seat in the agency's Executive Board, would lose leverage, would have less opportunity to influence development of the UNESCO programs that would be in effect in the medium-term future, and would lose a major role in selecting the next Director-General. The number of Americans in the

professional staff would also gradually decline.

This view was the opposite of that expressed by a Monitoring Panel appointed by Secretary of State Shultz to monitor the extent of change in the activities of and within UNESCO during 1984, especially in the subjects of greatest concern to the United States: that the interests of "minority groups of countries," which provide most of the financing, are inadequately protected; that UNESCO has deviated from its original purposes; that in solving problems it has adopted an excessively "statist" approach, as opposed to a "Western" approach; and that it is mismanaged, especially as regards its budget. This Monitoring Panel reported to the Secretary of State in late November that although there was considerable discussion and some progress, there was no "concrete" change.

The U.S. government stated its position and concerns regarding UNESCO's draft program and budget for 1986-87 in a July letter to Director-General M'Bow. It expressed desire for an increase in the authority of the member states by strengthening the General Conference and, in particular, the Executive Board; creation of a mechanism to ensure that major decision and programs have the support of all geographic groups, included the one that pays most of UNESCO's budget; return to concentration on UNESCO's core areas; changes in budget and management; and decentralization of power within the Secretariat.

Following discussion at the Commission's meeting of the proposed and evidently imminent actual U.S. withdrawal, one of the Commissioners circulated a proposed resolution "that it is not in the best interest of the United States to continue to participate and fund educational, scientific and cultural activities through membership in UNESCO," the Commission endorses President Reagan's decision to withdraw U.S. membership, fund-

ing, and support from UNESCO, and requests that the U.S. Congress disband the U.S. National Commission for UNESCO. This resolution was not acted upon. It gave way to a compromise resolution which stated that the U.S. notice of withdrawal had launched a process of reform in UNESCO, reaffirmed the Commission's conclusion that membership in UNESCO is in the U.S. national interest. It also recommended that if the United States withdraws, it should cooperate with the Director-General and other member states to encourage reforms that would permit the United States to rejoin the agency, and that the U.S. government involve representatives of the private sector in the effort to obtain such reforms. This resolution, with some amendments, passed by a vote of 37 (including your representative's vote) in favor to 9 against.

Withdrawal of the United States from UNESCO naturally would raise the question of the future of the U.S. National Commission. It was argued, on the one hand, that since the United States, in announcing that it would withdraw from UNESCO because of objections to its policies and administrative practices and procedures, expressed an intention to return to it if reforms removed those objections, continuation of the Commission

was desirable so it could advise the U.S. government as to whether reforms were sufficient, keep the private sector informed about UNESCO, and act as liaison between the private sector, UNESCO, and other national commissions. Further, if the United States later resumed membership in UNESCO, an advisory commission would again be needed and it would be better to retain the existing mechanism in some form than to abolish it and then have to recreate one. Against these arguments, it was argued that if the United States ceased to belong to UNESCO, the U.S. government did not need a body to advise it regarding its policies toward that agency; the Commission would cease to have grounds for continued existence. A resolution approved by the Executive Commission was offered stating that the Commission has an "especially important mission in the period ahead," and recommending that its officers and Executive Commission continue to press for official support, restoration of adequate funding and staff during "this transitional phase," and "remain seized of all questions related to the future of the National Commission." This resolution passed by a vote of 35 to 9 votes.

WALTER S. SALANT, *Representative*

## The Committee on the Status of Women in the Economics Profession

Women continue to increase their representation in the economics profession, but the rapid entry of young women occurring in other elite professions is not yet evident in economics. In 1983-84, the group of economics departments that grant most of the Ph.D.s (the so-called Chairs' Group) reported that 16 percent of their doctorates went to women. While this is an advance from the levels of the 1970's, it is below the proportion of women among newly trained lawyers (32 percent in 1980-81), physicians (25 percent), and chemical engineers (19 percent). Among undergraduates, the proportion of mathematics majors who are women (43 percent) exceeds the proportion of economics majors who are women (32 percent).

The economics profession continues to appear to undergraduate students as inhospitable to women. While the President-elect and one of the two Vice Presidents of the American Economic Association are currently women, the undergraduate is influenced by what she sees on her own campus. Surely a major factor in the perpetuation of this inhospitable image in the minds of today's students is the fact that many academic departments continue to be 100 percent male in their senior ranks. Some departments are 100 percent male in their entirety. In 1983-84, the situation with respect to the senior ranked positions was

	Number of departments	Number of women above rank of Asst. Prof.	At least this many departments with no women above Asst. Prof.:
Chairs' Group	41	22	19
Other Ph.D. granting depts.	34	21	13
Depts. granting MA only	46	27	19
Depts. granting BA only	189	49	140

The 41 departments of the Chairs' Group who reported on the composition of their faculties to the annual American Economic

Association survey, employed altogether 22 women as Full Professor or Associate Professor. We can deduce from this that at least 19, and surely more than half of them, had not a single woman above the rank of Assistant Professor.

Promotions for women within departments are less frequent than for men, given their representation in junior faculty positions (see Table 1). What is perhaps just as damaging is the fact that the ability of women to move from one school to a senior position in another school appears to be virtually nil. Of the 34 economists hired for senior positions by the departments of the Chairs' Group; only one was a woman. In all departments throughout the country, only two women made such a move.

In part to promote the visibility of women economists already in academic positions, CSWEP compiles and publishes a list of women faculty members at institutions which grant graduate degrees in economics. The women economists on that list should be prime candidates for recruitment by other academic departments. In the coming year, we will continue to update this list, so that it will be of greater use to economics departments who want to recruit women to their senior positions. Another project currently under examination for feasibility is the publication of a bibliography of women economists' scholarly publications, based on the *Journal of Economic Literature*.

We also plan to begin compiling lists of departments with no women faculty on senior levels or no women faculty at any level. In future years those lists should grow shorter and shorter, as more and more departments implement plans to end their exclusively male composition.

### Few Women Researchers Affiliated with the National Bureau

One of the most important functions of CSWEP is to campaign for the inclusion of

TABLE 1—DISTRIBUTION OF FULL-TIME FACULTY, BY TYPE OF INSTITUTION, ACADEMIC YEAR 1983-84

	Chair's Group			Other Ph.D.			Only M.A. Departments			Only B.A. Departments		
	Female		Total	Female		Total	Female		Total	Female		Total
	No.	Percent		No.	Percent		No.	Percent		No.	Percent	
Existing												
Professor	605	15	2.5	313	11	3.5	227	11	4.8	335	25	7.5
Associate	242	17	7.0	200	10	5.0	183	16	8.7	279	24	8.6
Assistant	315	32	10.2	205	29	14.1	158	27	17.1	346	57	16.5
Instructor	37	4	10.8	40	13	32.5	35	13	37.1	92	22	23.9
Other	39	9	23.1	21	1	4.8	27	13	48.1	40	1	2.5
New Hires												
Professor	24	1	4.2	6	0	0	0	0	0	4	0	0
Associate	10	0	0	4	0	0	3	0	0	3	1	33.3
Assistant	60	8	13.3	35	5	14.3	34	1	2.9	63	14	22.2
Instructor	19	2	10.5	14	2	14.3	14	3	21.4	42	12	28.6
Other	3	2	66.7	8	0	0	2	1	50.0	8	1	12.5
Promoted to Rank (1982-83)												
Professor	21	1	4.8	15	1	6.7	9	0	0	29	3	10.3
Associate	25	2	8.0	18	3	16.7	15	1	6.7	32	6	18.8
Assistant	1	0	0	4	2	50	1	0	0	14	2	14.3
Tenured at Rank (1982-83)												
Professor	1	0	0	9	0	0	1	0	0	8	2	25.0
Associate	12	2	16.7	28	3	10.7	15	1	6.7	22	5	22.7
Assistant	1	0	0	4	0	0	4	3	75.0	18	1	5.6
Other	0	0	0	0	0	0	0	0	0	0	0	0
Not Rehired												
Professor	29	0	0	20	1	5.0	9	1	11.1	7	0	0
Associate	9	0	0	11	1	9.1	5	1	20.0	6	1	16.7
Assistant	40	7	17.5	20	1	5.0	26	4	15.3	35	6	17.1
Instructor	10	2	20.0	10	2	20.0	10	3	30.0	19	3	15.7
Other	8	0	0	2	1	50.0	0	0	0	6	0	0

Note: In the tables for 1982-83 appearing in the May 1984 AEA *Proceedings*, numbers listed as referring to "Other Ph.D. Departments" actually refer to all Ph.D. departments.

women economists in all of the important activities in which professional economists are engaged. For almost a decade, the leadership of CSWEP has been particularly concerned with the situation at the National Bureau of Economic Research, where women have been largely excluded. On November 20, 1984, the present Chair of CSWEP and the two previous Chairs (Elizabeth Bailey, Dean, Graduate School of Industrial Administration, Carnegie-Mellon University, and Ann Friedlaender, Dean and Professor of Economics, Massachusetts Institute of Technology) signed a long letter to NBER President Martin Feldstein. In part, the letter said:

We at CSWEP are concerned about the low level of representation of wom-

en in the activities of the NBER. Currently only 6 of 170 Bureau research associates are women (2.8%), a level which has not shown any tendency to increase over the years since you became President. Yet the Bureau conducts research in a number of fields of applied economics in which women economists are active. We are concerned that the Bureau's low representation of women, combined with its steadily growing size and command over research funds, is increasingly putting younger women economists at a disadvantage relative to male economists in the same fields who have Bureau affiliations. We would like to urge you to take concrete measures to change this situation and we want to provide whatever help and guidance we can.

(Continued)

TABLE 2—PREVIOUS ACTIVITY OF NEW HIRES AND CURRENT ACTIVITY OF THOSE NOT REHIRED  
BY TYPE OF INSTITUTION AND SEX, ACADEMIC YEAR, 1983–84

	Previous Activity of New Hires				Current Activity of Not Rehiired			
	Male		Female		Male		Female	
	No.	Percent	No.	Percent	No.	Percent	No.	Percent
Chairs' Group	122	100.0	18	100.0	90	100.0	14	100.0
Faculty	39	32.0	3	16.7	61	67.8	6	42.9
Student	66	54.1	12	66.7	2	2.2	1	7.1
Government	1	.8	0	0	8	8.9	2	14.3
Bus., Banking, Research	1	.8	1	5.6	5	5.6	2	14.3
Other	15	12.3	2	11.1	14	15.6	3	21.4
Other Ph.D.	52	100.0	6	100.0	43	100.0	7	100.0
Faculty	15	28.8	1	16.7	25	58.1	3	42.9
Student	28	53.8	4	66.7	2	4.7	0	0
Government	2	3.8	0	0	5	11.6	2	28.6
Bus., Banking, Research	4	7.7	0	0	4	9.3	1	14.3
Other	3	5.8	1	16.7	7	16.3	1	14.3
M.A. Departments	52	100.0	6	100.0	30	100.0	8	100.0
Faculty	15	28.8	0	0	17	56.7	4	50.0
Student	33	63.5	5	83.3	2	6.7	2	25.0
Government	0	0	0	0	0	0	0	0
Bus., Banking, Research	0	0	0	0	5	16.7	0	0
Other	4	7.7	1	16.7	6	20.0	2	25.0
B.A. Departments	127	100.0	37	100.0	66	100.0	14	100.0
Faculty	36	28.3	10	27.0	30	45.5	8	57.1
Student	75	59.1	18	48.6	12	18.2	1	7.1
Government	3	2.4	0	0	1	1.5	0	0
Bus., Banking, Research	10	7.9	4	10.8	14	21.2	0	0
Other	3	2.4	5	13.5	9	13.6	5	35.7

Note: See Table 1.

TABLE 3—DISTRIBUTION OF SALARY FOR WOMEN FACULTY BY TYPE OF DEPARTMENT AND TIME IN RANK,  
ACADEMIC YEAR, 1983–84

Relative Salary for Rank	All Women		Time in Rank			
	Number	Percent	Total Percent	Above Median	At Median	Below Median
All Departments	309	100.0	100.0	29.1	42.7	28.2
Salary above Median	105	34.0	100.0	57.1	25.7	17.1
Salary at Median	109	35.3	100.0	15.6	77.1	7.3
Salary below Median	95	30.7	100.0	13.7	22.1	64.2
Ph.D., Chair's Group	70	100.0	100.0	28.6	31.4	40.0
Salary above Median	22	31.4	100.0	40.9	31.8	27.3
Salary at Median	17	24.3	100.0	41.2	35.3	23.5
Salary below Median	31	44.3	100.0	12.9	29.0	58.1
Ph.D., Other	58	100.0	100.0	36.2	36.2	27.6
Salary above Median	24	41.4	100.0	70.8	20.8	8.3
Salary at Median	14	24.1	100.0	7.1	71.4	21.4
Salary below Median	20	34.5	100.0	15.0	30.0	55.0
M.A. Departments	74	100.0	100.0	33.8	39.2	27.0
Salary above Median	20	27.0	100.0	80.0	15.0	5.0
Salary at Median	27	36.5	100.0	18.5	81.5	0
Salary below Median	27	36.5	100.0	14.8	14.8	70.4
B.A. Departments	107	100.0	100.0	22.4	56.1	21.5
Salary above Median	39	36.4	100.0	46.2	30.8	23.1
Salary at Median	51	47.7	100.0	7.8	90.2	2.0
Salary below Median	17	15.9	100.0	11.8	11.8	76.5

Note: See Table 1.



TABLE 4—DEGREES GRANTED IN ECONOMICS BY TYPE OF DEPARTMENT AND SEX, ACADEMIC YEAR, 1983–84

Number of:	All Depts.	Ph.D. Departments			M.A. Depts.	B.A. Depts.
		Total	Chairs'	Other		
Departments	377	120	44	76	45	212
Ph.D.s	542	542	406	136	—	—
Female	86	86	66	20	—	—
Percent Female	15.9	15.9	16.3	14.7	—	—
M.A.s	1229	1000	639	361	229	—
Female	279	236	135	101	43	—
Percent Female	22.7	23.6	21.1	28.0	18.8	—
B.A.s	12285	7292	5058	2234	1006	3987
Female	3912	2214	1554	660	270	1428
Percent Female	31.8	30.4	30.7	29.5	26.8	35.8
Other	127	53	33	20	23	51
Female	39	13	7	6	6	20
Percent Female	30.7	24.5	21.2	30.0	26.1	39.2

Note: See Table 1.

Obviously, membership in one of the NBER's permanent research programs provides tremendous benefits to younger academics.... Despite the substantial benefit from belonging to one of these programs, there appears to be no formal selection procedure that would ensure that the best researchers in each field are represented. Most research associates/fellows appear to be either former students of directors or senior research associates of that group or junior faculty at a few leading universities. Apparently no attempt is made to publicize these positions or to allow outsiders to apply. Due to the extent that women are not part of the "old-boy network" linking the Bureau research associates, they are effectively eliminated from the pool of potential associates....

What can be done to remedy this situation? We have several suggestions.

First, many male economists were brought into Bureau association because they were either Ph.D. students of Bureau project directors (or other Bureau research associates) or were junior faculty members in the University departments where Bureau project directors teach. Therefore, one way that we advocate for bringing in more women researchers is for all NBER project directors and research associates to review their Ph.D. students of the last

five years and the recent women hires in their departments and to consider bringing in any interested and qualified women economists who have been passed over. If, on the other hand, on doing this they find that they have had no women Ph.D. students over the past five years, then perhaps they should ask themselves why not and consider seriously whether they have been practicing unconscious sex discrimination in selection of thesis students....

Second, some procedure should be set up to allow "outsiders" to apply for positions as NBER research fellows in each group. Since the group of research associates/research fellows is by invitation only, it provides little opportunity for women to gain entry, since the NBER's project directors and other senior researchers have been very unlikely in the past to bring them in. This means that women economists are likely to be excluded by virtue of the selection process even if they are part of the pool of distinguished economists working in areas of interest to the Bureau.

We await your reply and, again, offer our help and guidance as you consider what concrete measures would be best adopted to rectify this situation.

Sincerely,

...

TABLE 5—DISTRIBUTION OF ACTIVITIES OF NEW PH.D. DEGREES BY SEX AND TYPE OF DEPARTMENT, ACADEMIC YEAR, 1983-84

	All Ph.D. Depts.		Chairs' Group		Other Ph.D. Depts.	
	No.	Percent	No.	Percent	No.	Percent
All Ph.D.s	468	100.0	371	100.0	97	100.0
Education	279	59.6	224	60.4	55	56.7
Government	42	9.0	32	8.6	10	10.3
Bus., Banking, Research	33	7.1	27	7.3	6	6.2
Int'l Emp. Outside U.S.	65	13.9	47	12.7	18	18.6
Other	49	10.5	41	11.1	8	8.2
Male Ph.D.s	377	100.0	292	100.0	85	100.0
Education	229	60.7	181	62.0	48	56.5
Government	31	8.2	22	7.5	9	10.6
Bus., Banking, Research	26	6.9	20	6.8	6	7.1
Int'l Emp. Outside U.S.	55	14.6	38	13.0	17	20.0
Other	36	9.5	31	10.6	5	5.9
Female Ph.D.s	91	100.0	79	100.0	12	100.0
Education	50	54.9	43	54.5	7	58.3
Government	11	12.1	10	12.7	1	8.3
Bus., Banking, Research	7	7.7	7	8.9	0	0
Int'l Emp. Outside U.S.	10	11.0	9	11.4	1	8.3
Other	13	14.3	10	12.7	3	25.0

Note: See Table 1.

TABLE 6—DISTRIBUTION OF PH.D. STUDENT SUPPORT, BY TYPE OF SUPPORT, SEX, AND DEPARTMENT, ACADEMIC YEAR, 1983-84

	All Ph.D. Depts.		Chairs' Group		Other Ph.D. Depts.	
	No.	Percent	No.	Percent	No.	Percent
All Students	3973	100.0	3099	100.0	874	100.0
Tuition Only	183	4.6	140	4.5	43	4.9
Stipend Only	365	9.2	215	6.9	150	17.2
Tuition + Stipend	1972	49.6	1589	51.3	383	43.8
No Support	1230	31.0	1003	32.4	227	26.0
No Record	223	5.6	152	4.9	71	8.1
Male Students	3118	100.0	2455	100.0	663	100.0
Tuition Only	145	4.7	111	4.5	34	5.1
Stipend Only	299	9.6	184	7.5	115	17.3
Tuition + Stipend	1505	48.3	1228	50.0	277	41.8
No Support	979	31.4	787	32.1	192	29.0
No Record	190	6.1	145	5.9	45	6.8
Female Students	855	100.0	644	100.0	211	100.0
Tuition Only	38	4.4	29	4.5	9	4.3
Stipend Only	66	7.7	31	4.8	35	16.6
Tuition + Stipend	467	54.6	361	56.1	106	50.2
No Support	251	29.4	216	33.5	35	16.6
No Record	33	3.9	7	1.1	26	12.3

Note: See Table 1.

### Representation at Annual Meetings

Any process of professional selection that is informal, and whose details are only known or understood by a relatively small in-group,

are disadvantageous to women, who benefit less frequently than men from sponsorship by more established members of the profession. The process by which sessions at the AEA annual meetings are organized and

papers invited has been one of these little-understood processes. Formally speaking, the President-elect does the inviting; in practice, many volunteers communicate to him or her their desire to participate, and it is out of these submissions that a considerable part of the program is in fact assembled with the help of referees.

We at CSWEP will continue to urge that the selection procedures for the annual meetings be made more formal and more public. While there is an understandable interest in having the profession's (mostly male) celebrities on parade at the meetings, we would urge procedures which give a better representation to innovative research from the less well-connected members of the profession, women among them. In the meantime, through our *Newsletter* we are urging women economists to submit proposals for sessions or individual papers to the President-elect.

#### Research on Gender-Related Topics

CSWEP has been concerned to encourage and foster research on gender roles in the economy and related policy issues, and to make sure that women economists and points of view sensitive to the special problems many women face under current economic institutions are well represented in the field.

To this end, we continue to sponsor sessions on these topics at the AEA and regional meetings. In November 1984, CSWEP jointly with The Brookings Institution sponsored a conference on Gender Issues in the Workplace, arranged by Clair Brown and Joseph Pechman.

As research proceeds and interest rises, courses on the economics of gender roles are being offered at an increasing number of schools. At least three new textbooks are in the works. A number of economics departments are specifically looking to recruit a specialist in gender-related topics, and individual economists are "coming out" as specialists in the field.

#### Committee Operation

We wish to thank Gail Wilensky and Nancy Ruggles, whose terms on the Committee expires this year. Gratitude is also due to Aleta Styers, who continues to bear the time-consuming editorial duties on the *Newsletter* with relative fortitude, and to Joan Haworth, who served as Membership Secretary. New Committee members for 1985 are Helen Junz of the IMF and Karen Davis of the Johns Hopkins School of Public Health.

BARBARA R. BERGMANN, *Chair*

## Report of the Committee on U.S.–China Exchanges

As the reports of our Committee in the last three years reveal, there has been continued and expanding exchange activities between American and Chinese economists. In 1984, a great step was taken by the Chinese government to make post-Marxian modern economics official, in the sense of introducing it to the Chinese university curriculum, side by side with Marxian economics, as being important for China's economic development. Previously, modern economics was taught as an elective for the understanding of capitalist economies only.

To illustrate the change in viewpoint, the Chinese Ministry of Education sponsored a workshop in microeconomics for university teachers, graduate students, and government officials. The workshop was held at Peking University from June 11 to July 21. The lecturers were Sherwin Rosen (micro theory, human capital and labor), Marc Nerlove (micro theory, agricultural supply, and empirical studies), Edwin Mills (development, urban, and environmental economics), and Gregory Chow (applications to China). Microeconomics was to become an important part of economics education in China. Workshops in macroeconomics and econometrics are planned for the summers of 1985 and 1986, respectively.

The Ministry of Education decided to send students abroad to study economics. On November 17–20, TOEFL, mathematics, and economics examinations (given in English, with economics questions provided by Mills, Rosen, and Chow) were administered by the Ministry to 151 candidates who had been carefully selected from twenty-seven major Chinese universities and economics and finance colleges, using scholastic records and preliminary tests. Eighty-one candidates, who received a total grade of 120 or higher in mathematics and economics, have been rec-

ommended for graduate work to sixty-one American and Canadian universities, including all major economics departments.

Both the Ministry of Education and the Chinese Academy of Social Sciences (CASS) are seeking financial support to strengthen economics education and research at the universities and economic research institutes. To consider such support, the Ford Foundation has formed a committee, co-chaired by Dwight Perkins and Gregory Chow. Activities to be supported might include workshops, Boulder-like institutes, and research programs in China, fellowships for students and visiting scholars from and to China, joint research projects, and library and computer facilities.

In 1984, many American economists lectured in China. Many Chinese economists visited the United States, including a delegation from CASS in September, mentioned in our Report last year. From 1984 on, many teaching and research opportunities in China are becoming available to American economists. To find a suitable hosting institution, a potential visitor can consult published material such as bulletins of universities and research institutes, and Chinese economics journals reporting on research conducted at various institutions. Helpful information can be found in Teh-wei Hu's research report, "The State of American Economic Studies in the People's Republic of China" (U.S. Information Agency: April 1984) and in Gregory Chow's report, "Economic Research in China" (Princeton University, 1984). Personal contacts will also be useful to generate a suitable visiting arrangement, as in the case of searching for a visiting position in the United States.

GREGORY C. CHOW, *Chair*

## Report of the Committee on Economic Education

The Committee this year (1984) closed out its work on the economics major project which had been directed by John Siegfried, carried out in conjunction with the Joint Council on Economic Education, and financed by the Alfred P. Sloan Foundation. The project resulted in a series of publications that provide benchmark data on the characteristics of economics major programs at the college and university level, and of economics major students enrolled in a subset of these institutions. The data collected in the study are available for further analysis (for information, contact John Sumansky at the Joint Council on Economic Education, 2 Park Avenue, New York, NY 10016).

The Committee is exploring the possibility of convening a small conference to follow up on the study of the economics major. This conference would be concerned with trying to specify more fully what it is that we would like economics majors to be able to do upon graduation. What knowledge and skills might we reasonably expect them to possess? What competencies might we expect them to be able to demonstrate? No unanimity of views about the answers to these questions is anticipated. Nevertheless, it is hoped that the conference and resulting papers would aid departments and sharpen their understanding of what they are attempting to do in educating their majors.

The Committee has also reviewed its principal on-going project, the Teacher Training Program, carried out in collaboration with the Joint Council in Economic Education. The Program, initiated in 1972 with the help of a grant from the Alfred P. Sloan Foundation, led to the institutionalization of teacher training programs for teach-

ing assistants, and in some cases for faculty members at a number of major universities; it also produced a Resource Manual and accompanying video tapes to support these programs. The second phase of this project, supported by the Lilly Endowment Inc., led to the development of an annual one-week seminar designed to help improve the teaching skills of economics faculty members and to help their departments establish their own teacher training programs. The third phase, supported by Citibank and Borg-Warner, funded regional seminars whose purpose was to build a cadre of trainers that would assist still other departments establish their own teacher training programs.

The Committee decided to undertake a thorough evaluation of the TTP program, prepare a new edition of the Resource Manual, and develop a new set of videotapes. The evaluation will be done by Mike Salemi and should be completed within the year. Work on revising the Resource Manual will begin later this year, followed by preparation of the videotapes.

The Committee continues its efforts to build the circulation and quality of the *Journal of Economic Education*. Just a year ago the *Journal's* focus was greatly broadened. Traditionally, it reflected an almost exclusive emphasis on research articles dealing with economic education. Since then its scope has been expanded to include content articles on economics, papers on the teaching of economics, innovations in economic instruction, and information of general interest to the profession.

W. LEE HANSEN, *Chair*

## Report of the *Ad Hoc* Committee on Financial Reporting and Changing Prices

The *Ad Hoc* Committee was set up by the President of the AEA at the request of the Financial Accounting Standards Board to help it reach a decision as to whether the inflation disclosures mandated five years earlier through Statement No. 33, on an experimental basis, should be continued in some form, permanently or experimentally, or should be discontinued.

The Committee submitted a unanimous preliminary report in August 1984 and a final report on December 12, 1984. The full report (copies can be secured upon request to the Chair) reviews a substantial body of evidence on whether the inflation disclosures have provided "valuable" information. It concludes that the evidence is rather negative, both with respect to the use of the information by various market participants and in terms of the association between the information disclosed and the market valuation of equity. It finds, however, that it is not possible from this evidence to establish whether the negative findings reflect: (i) the ability of the market to secure the information from sources other than, and in advance of, the inflation disclosure; or (ii) the poor quality and lack of standardization of the information disclosed; or, finally, (iii) the fact that potential users (wrongly) regard the information as intrinsically worthless.

The report stresses that an understanding of the main reason for the prevailing disre-

gard of the information disclosed under Statement 33 is crucial in deciding whether it is worth continuing such disclosures. Accordingly, it recommends a number of steps designed to probe further into the reason for the puzzling, apparent failure on the part of a variety of potential users to utilize the information, and create conditions for increased future utilization. It further recommends that, while these steps are being implemented, the requirement of some form of inflation disclosures be continued for an experimental period of up to five years.

With respect to revisions in the content of the current disclosures, the Committee submitted two main suggestions. The first relates to the measurement of capital consumption and recommends not dropping the financial capital maintenance concept—as implemented through a general price-level adjustment on long-term assets and shareholders' equity—in favor of a replacement cost (physical capital maintenance) concept. The second recommendation, relating to the inflation adjustment of nominal interest cost, is that the adjustment should be applied not to the book values of assets, as currently required, but rather to the market value, and that it should also include the change, if any, in the market valuation of the debt.

FRANCO MODIGLIANI, *Chair*

# Essential Reading for the Economist

## **The U.S. Economy Demystified**

*What the Major Economic Statistics Mean and Their Significance for Business*

Albert T. Sommers, The Conference Board

Foreword by James T. Mills

"Al Sommers is one of the very few economists in this country who understands the subtle relationships which exist between our data system and the complex economy which it is endeavoring to measure."—Alan Greenspan

"Every academic, business, and government economist concerned with the course of the economy from month to month and year to year has learned to respect the insights of Al Sommers. He knows the numbers thoroughly and interprets them with perception and wisdom, based on his own understanding of the mechanisms at work, born of long experience."—James Tobin

April ISBN 0-669-09427-7 \$16.95

In paper: 0-669-09821-3 \$10.00

## **The Global Financial Structure in Transition**

*Consequences for International Finance and Trade*

Joel McClellan, the Quaker United Nations Program, editor

This book demonstrates the wide diversity of opinions on how to understand and manage the rapidly changing forces in international finance and trade. Topics include international indebtedness, trade deficits, long-term investments, and refinancing and trade expansion.

May ISBN 0-669-09581-8 ca. \$32.00

## **Political Economy and Risk in World Financial Markets**

Tamir Agmon, Tel-Aviv University

The author, an internationally known Israeli economist, presents specific strategies under two broad headings, risk avoidance and risk negotiation. He illustrates these with case studies and a mathematical model describing the changing relations between a multinational corporation and the host government of a newly industrialized country.

June ISBN 0-669-08339-9 ca. \$22.50

## **The Political Economy of Coal**

Ferdinand E. Banks, University of Uppsala and the University of Melbourne

An encyclopedic primer on worldwide energy economics, with a particular emphasis on coal. The author looks at coal production's effect on the world economy and considers how various countries' reserves affect the balance of power.

288 pages ISBN 0-669-06109-7 \$29.00

125 Spring Street

Lexington, MA 02173

(617) 862-6650 (212) 924-6460



**Lexington Books**

# *Decision Analysis Takes a Giant Step Down*

Accurate planning and decision making no longer demand expensive main-frame and connect charges. The process requirements have been reduced to:

- IBM PC (minimum 256K)
- DOS 2.0
- OTIS

**OTIS** is a decision-support command language designed to manage time series data for analysis and forecasting. Intuitively easy to use and understand, the system combines into a single program the following features:

*Time series analysis*

*Data and file management*

*Graphics and data display*

*Report generating*

*Command processing*

*Statistical analysis*

*Single equation solving*

*Menu processing*

*Regression analysis*

*Correlation analysis*

*Simultaneous equation  
simulations*

Although stand-alone, **OTIS** is compatible with other popular software programs, such as Lotus 123, Symphony, and dBase.

Time series are developed and managed in a "workspace" (memory) environment. Series can be created and transformed using the usual arithmetic operations, and functions such as square root, logarithm, and exponential. In addition, observations can easily be modified, added to, and deleted from new and existing series. Each series contains a comment to describe its purpose and/or creation. Series may be displayed as listing, character plots and/or line graphs. Equations can be generated for purposes of solving singly or simultaneously. Series, equations, and parameters can be stored and retrieved in disk files which are password protected. Moreover, series can be transferred easily (in ASCII format) between **OTIS** and other popular software programs, such as Lotus 123, Symphony, dBase and Framework.

The program is command driven for analytical speed with "help" reminders available at all times. Ease of use is enhanced by requiring for execution only the first two characters of any command. To facilitate customized applications, **OTIS** also provides a command programming language and menu processor. Finally, commands can be stored and executed from an executable disk file.

demo diskette    \$10  
manual            \$25  
OTIS              \$995



**ODIN RESEARCH**  
834 Old State  
Berwyn, PA 19312  
(215) 296-4485

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*





New titles from  
**The National Bureau of Economic Research:**

**The Structure  
and Evolution of  
Recent U.S. Trade Policy**

*Edited by Robert E. Baldwin and  
Anne O. Krueger*

This volume represents the first systematic effort to analyze specific U.S. trade policies, particularly nontariff measures. It provides a better understanding of how trade policies operate, of how effective they are, and of their costs and benefits to trading nations.

**An NBER Conference Report**  
Cloth \$50.00 448 pages

**Exchange Rate  
Theory and Practice**

*Edited by John F. O. Bilson and  
Richard C. Marston*

Here the world's most respected international monetary economists discuss three significant new approaches to the economics of exchange rates, test and question these approaches with evidence from empirical studies, and analyze a number of exchange rate policies in use today, including the European Monetary System.

**An NBER Conference Report**  
Cloth \$58.00 544 pages

**A Retrospective on the  
Classical Gold Standard,  
1821–1931**

*Edited by Michael D. Bordo and  
Anna J. Schwartz*

The studies in this volume were designed to gain a better understanding of the historical gold standard, but they also throw light on the question of whether restoring it today could help cure inflation, high interest rates, and low productivity growth.

**An NBER Conference Report**  
Cloth \$65.00 694 pages

**Economic Transfers  
in the United States**

*Edited by Marilyn Moon*

In recent years the definition of an economic transfer — a payment to an individual or institution that does not arise out of current productive activity — has been subject to ever wider interpretation. This volume addresses that trend and introduces new methods of measuring transfers in the American economy. Researchers, policy analysts, and those who compile statistics on which social programs are based will value the diverse approaches of these ten papers and their accompanying comments.

**NBER Studies in Income and  
Wealth, Volume 49**  
Cloth \$42.00 400 pages

The University of **CHICAGO** Press

5801 South Ellis Avenue, Chicago, IL 60637

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

*The real world  
comes to the classroom. . .*

**Robert J. Barro**  
*University of Rochester*  
**MACROECONOMICS**

Barro's Macroeconomics is the first modern presentation of the market clearing approach to macroeconomics. It provides an honest and realistic alternative to the cumbersome and unrewarding IS/LM approach of current intermediate texts.

"Robert Barro has been a leading contributor to recent developments in the analysis of the overall behavior of the economy. He has now written a lucid, comprehensive, and authoritative exposition of the current state of our understanding of these phenomena. His text covers both the latest theoretical developments and the latest empirical evidence. It can be highly recommended on both levels."

—Milton Friedman,  
Senior Research Fellow  
Hoover Institution on War, Revolution and Peace,  
Stanford University

"... What Barro does is not only build a macroeconomic model from the ground up but also one that is meant to explain observed economic phenomena... if only other textbook authors would try harder to do this, students would learn more and teachers would find their job much easier."

—James Barth,  
The George Washington University

1984     580 pp.

You owe it to yourself and your students to examine this important new text. To request a complimentary copy, write to Lisa Berger, Dept. 5-1926. Please include course name, enrollment, and title of present text.

**JOHN WILEY & SONS, Inc.**  
**605 Third Avenue, New York, N.Y. 10158**

In Canada 22 Worcester Road, Rexdale, Ontario M9W 1L1



5-1926

*Keeping pace with the latest  
developments in economics*

---

**THE THEORY AND  
PRACTICE OF ECONOMETRICS**  
SECOND EDITION

**George G. Judge**, *University of Illinois*

**William E. Griffiths**, *University of New England*

**R. Carter Hill**, *University of Georgia*

**Helmut Lütkepohl**, *Universität Osnabrück*

**Tsoungh-Chao Lee**, *University of Connecticut*

Now the most complete treatment of major econometric problems has been revised and updated. The authors present the most systematic and up-to-date view of econometric problems, their statistical consequences, remedies, alternatives, and future research. Includes new chapters on asymptotic distribution theory, Bayesian inference, time series, and simultaneous equation statistical models. The distribution lag chapters have been rewritten to tie-in with time-series chapters.

Due January 1985    approx 1050 pp.    ISBN 0-471-89530-X

---

*Other New Titles:*

**AGRICULTURAL ECONOMICS AND AGRIBUSINESS**, Third Edition

**Gail L. Cramer**, *Montana State University*

**Clarence W. Jensen**, *Montana State University*

January 1985    approx 475 pp.    ISBN 0-471-87871-5

**ECONOMICS OF PRODUCTION**

**Bruce R. Beattie**, *Montana State University*

**C. Robert Taylor**, *Montana State University*

January 1985    approx 320 pp.    ISBN 0-471-80810-5

---

*1984 Titles of Interest:*

**AMERICAN MONEY AND BANKING**

**Maxwell J. Fry**, *University of California*

**Raburn M. Williams**, *University of Hawaii*

**MONETARY AND FINANCIAL ECONOMICS**

**James L. Pierce**, *University of California, Berkeley*

**URBAN LAND ECONOMICS**

**Michael A. Goldberg**, *University of British Columbia*

**CASES IN MANAGERIAL ECONOMICS**, Second Edition

**Bernard J. Winger**, *University of Dayton*

To request a complimentary copy, write to Lisa Berger, Dept. 5-1926. Please include course name, enrollment, and title of present text.

**JOHN WILEY & SONS, Inc.**

**605 Third Avenue, New York, N.Y. 10158**

In Canada 22 Worcester Road, Rexdale, Ontario M9W 1L1



5-1926

# AEA sponsored Group Life Insurance for you and your family— at attractive rates!

The AEA Group Life Insurance Plan can help provide valuable supplementary protection—at attractive rates—for eligible members and their dependents.

Because AEA participates in a large Insurance Trust which includes other scientific and technical organizations, the low cost may be even further reduced by premium credits. In the past eight years, insured members received credits on their April 1 semiannual payment notices averaging over 44% of their annual premium contributions. (These credits are based on the amount paid during the previous policy year ending September 30.) Of course future premium credits, and their amounts, cannot be promised or guaranteed.

Now may be a good time for you to re-evaluate your present coverage and look into AEA Life Insurance. Just fill out and return the coupon for more details at no obligation.

<b>Administrator, AEA Group Insurance Program</b>		<b>G-3</b>
<b>1255 23rd Street, N.W.</b>		
<b>Washington, D.C. 20037</b>		
Please send me more information about the AEA Life Insurance Plan.		
Name _____	Age _____	
Address _____		
City _____	State _____	Zip _____

Or—call today Toll-Free 800-424-9883  
(Washington, DC area, call 296-8030)

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

## **Deregulating the Airlines**

*Elizabeth E. Baily, David R. Graham, and Daniel P. Kaplan*

For policymakers and students of regulation in particular, this study provides a unique case for contrasting the operation of an industry under close regulatory control and its operation free of such controls. It is able to make use of an unusually large volume of data on the cost, operations, and prices of individual firms to show how markets work and how regulation works.

\$19.95

## **Profit Cycles, Oligopoly, and Regional Development**

*Ann Markusen*

Markusen provides a pioneering approach that will enable planners and managers to better cope with baffling changes in the current economic viability of regions. Her "profit-cycle theory" provides a key to understanding how, why, and when a region's leading industries undergo major changes.

\$25.00

## **Shifting Gears**

*Changing Labor Relations in the U.S. Automobile Industry*

*Harry Katz*

This book stands apart from other discussions of labor relations and American management weaknesses by combining historical, case, and statistical analyses to look at past and contemporary events, and to indicate the likely future course of auto bargaining.

\$22.50

*Now available in paperback*

## **Game Theory in the Social Sciences**

*Concepts and Solutions*

*Martin Shubik*

\$11.95

## **A Game-Theoretic Approach to Political Economy**

*Volume 2 of Game Theory in the Social Sciences*

*Martin Shubik*

\$47.50

## **Retirement, Pensions, and Social Security**

*Gary S. Fields and Olivia S. Mitchell*

"A careful econometric analysis of actual and potential changes in retirement policy."—Steven H. Sandell, Director Project on National Employment Policy and Older Americans

\$19.95

*Original in Paperback*

## **Challenges and Choices Facing American Labor**

*edited by Thomas A. Kochan*

After decades of stability, labor-management relations are undergoing dramatic changes. This collection provides the best and most current summary of the extent and causes of today's upheaval in industrial relations.

\$15.00 (cloth \$30.00)

28 Carleton Street  
Cambridge, MA 02142

**THE MIT PRESS**

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

# Dividends on thought.

from Columbia and NYU



## Competition Among New Video Media

*Eli Noam.* This unique study assembles a series of works by academics, policy makers, and industry leaders to analyze the communication revolution and the impact of technology on regulation. *Columbia Studies in Business, Government, and Society, Eli Noam, General Editor.* 416 pp., illus., \$40.00 (June)

## Corporate Trust Administration and Management

Third Edition

*Robert I. Landau.* Completely updated to cover new laws and recent court decisions, this new edition is a basic reference on practices and emerging trends in trustee and agency activities and responsibilities. 448 pp., \$35.00

## The Logic of Collective Choice

*Thomas Schwartz.* An abstract, deductive theory of some of the more salient types of collective choice processes. 288 pp., \$30.00

## Economic Security and the Origins of the Cold War, 1945-1950

*Robert A. Pollard.* "... likely to become the definitive post-revisionist account."—John Lewis Gaddis. *The Political Economy of International Change, John Gerard Ruggie, General Editor.* 448 pp., \$30.00 (August)

## Shaky Palaces

Home Ownership and Social Mobility in Boston's Suburbanization

*Matthew Edel, Elliot D. Sclar, and Daniel Luria.* *Columbia History of Urban Life Series, Kenneth Jackson, General Editor.* 440 pp., A King's Crown Paperback. \$15.00; \$35.00 cl

## Land, Labor, and Rural Poverty

Essays in Development Economics

*Pranab K. Bardhan.* 288 pp., \$30.00

New from New York University Press:

## Working Together

Employee Participation in Action

*John Simmons and William Mares.* "A thorough, often insightful, overview of both successful and unsuccessful efforts at making work more cooperative and more democratic."

—*The New York Times.* 336 pp., Now in paperback \$11.50 pa

## Setting Municipal Priorities

American Cities and the New York Experience

*Charles Brecher and Raymond D. Horton, Editors.*

"... offers a wealth of information and analysis."—Nathan Glazer, *The New York Times.* 560 pp., \$20.00 pa, \$50.00 cl

## Unequal Treatment

A Study in the Neoclassical Theory of Discrimination

*Mats Lundahl and Eskil Wadensjö.* Applies the theory of economic discrimination—and the phenomenon of "crowding"—to analyze South Africa's apartheid policy. 336 pp., \$45.00

## Mathematical Methods in Economics

*Norman Schofield.* 296 pp., \$39.50

## The World Banking System

Outlook in a Context of Crisis

*Andrew F. Brimmer and Robert G. Hawkins.* *The Joseph I. Lubin Memorial Lecture Series, 1.* 60 pp., \$12.50

## The Multinational Banking Industry

*Neil Coulbeck.* Presents detailed country-by-country studies, case studies of banks, and over 100 tables and figures. 397 pp., \$55.00

Send check or money order to Dept. JN at the address below, including \$2.00 per order for postage and handling.

# Columbia University Press

136 South Broadway, Irvington, NY 10533

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

NEW BOOKS-NEW BOOKS-NEW BOOKS-NEW BOOKS-NEW

## ix

# Centennial T-Shirt



celebrating the  
100th Anniversary  
of the

**American  
Economic Association**  
1885 - 1985.

Now available by mail.

Your next chance: the year 2085

- Collector's Item
- Limited Edition
- Satisfaction Guaranteed

Top quality name brand T-Shirt (50% cotton and 50% polyester for easy care and no shrinkage) imprinted with the familiar AER cover design in authentic burgundy red with the AEA Seal in black, printed over your choice of light blue or tan fabric.

PLEASE PRINT CLEARLY

Send to:  
**ACADEMICS**  
1800 E. Capitol Drive  
Suite 2F  
P.O. Box 11768  
Milwaukee, WI 53211

Send \_\_\_\_\_ AEA's Centennial T-Shirts at \$9.95 each including handling and delivery.

Enclosed: check \_\_\_\_\_ money order \_\_\_\_\_ payable to ACADEMICS.

Or charge my MasterCard \_\_\_\_\_ VISA \_\_\_\_\_.

Card # \_\_\_\_\_ Exp. Date \_\_\_\_\_

Size Chart	
Small	34-36
Medium	38-40
Large	42-44
X-Large	46-48

Name \_\_\_\_\_

Address \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_ Zip \_\_\_\_\_

Signature \_\_\_\_\_

Quantity	Size	Color

Wisconsin residents add 5% sales tax. Allow 2-3 weeks for delivery.

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers



# Software for Hard Problems.

If you've outgrown your spreadsheet, if you're tired of waiting for the mainframe, consider STATA.<sup>™</sup>

- Like a spreadsheet program, STATA<sup>™</sup> is interactive. STATA<sup>™</sup> makes it easy to answer the "what if" questions that spreadsheet programs handle so well.
- Like a database management program, STATA<sup>™</sup> allows you to create complex data sets, transform them in almost any way you desire, make tables, and locate particular pieces of information.
- Like a statistical package, STATA<sup>™</sup> performs a variety of statistical analyses including multivariate regression.
- Like a programming language, STATA<sup>™</sup> allows you to store groups of STATA<sup>™</sup> commands so that complicated analyses can be automated.

And when you think your problem has outgrown your PC, think again. STATA<sup>™</sup> is the software for hard problems.

STATA<sup>™</sup> efficiently sorts, appends, and merges datasets. STATA<sup>™</sup> understands many different data formats (for example, STATA<sup>™</sup> can read Census data directly). STATA<sup>™</sup> can also format data for use by other programs. STATA<sup>™</sup> can encrypt data to keep private data private.

STATA<sup>™</sup> calculates all the standard univariate statistics, correlations and covariances, one-, two-, and n-way tables with chi-square tests for independence. STATA<sup>™</sup> estimates multivariate regression and ANOVA models on an unlimited

number of observations using ordinary least squares, instrumental variables, or two-stage least squares. STATA<sup>™</sup> also performs tests of linear hypotheses about these models. You can perform simulations and Monte Carlo studies using STATA's prediction and random number functions.

STATA<sup>™</sup> labels your variables and their values so that your output is as close to publication-ready as possible. STATA<sup>™</sup> makes tables and plots and can copy its output to a file that can be edited with your word processor.

All of this is available for the IBM PC, PC/XT, or PC/AT with 256K of memory and one double-sided disk drive, for \$395.

A demonstration diskette is \$15.00, and the payment can be applied to subsequent purchase. For more information, call us at (213) 470-4341 or write to us at: Computing Resource Center, 10801 National Boulevard, Los Angeles, California 90064. Academic and quantity discounts available.

---

## STATA<sup>™</sup>

---

The Data Tool.<sup>™</sup>

STATISTICS/DATA ANALYSIS

IBM is registered trademark of International Business Machines Corporation. STATA<sup>™</sup> is a trademark of Computing Resource Center. ©1985

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

# **The Journal of International Economics and Economic Integration Offers**

## **\$5,000**

### **For the First Annual Daeyang Prize in Economics**

- The Journal of International Economics and Economic Integration is published biannually by the Institute for International Economics, King Sejong University, Seoul, Korea.
- The purpose of the Journal of International Economics and Economic Integration is to support and encourage research in the area of international trade, international finance and other related economic issues that include general professional interest in international economic affairs.
- The Journal of International Economics and Economic Integration welcomes unsolicited manuscripts, which will be considered for publication by the Editorial Board.
- The Editorial Board will choose fourteen manuscripts for publication on an annual basis.
- The Editorial Board will choose the best manuscript out of the fourteen to be awarded the \$5,000 Daeyang Prize in Economics.
- The manuscripts, which should be accompanied by an abstract of no more than 100 words, should be typewritten, double-spaced, in English with footnotes, references, figures, tables and any other illustrative material on separate sheets.
- Three copies of the manuscript and all accompanying material should be submitted to the following address by October 31, 1985 for consideration for the 1986 publication.

**Institute for International Economics  
King Sejong University  
Seongdong-Ku, Seoul, Korea**



## ***Economics of Worldwide Stagflation***

*Michael Bruno and Jeffrey Sachs*

"This is *serious* supply-side economics . . . An excellent and important book."—Robert M. Solow

\$25.00

## ***Racial Conflict and Economic Development***

*W. Arthur Lewis*

Lewis discerns the ways in which race and economics affect individuals and groups, bringing a personal viewpoint to the problems faced by both less-developed and more-developed countries.

*The W. E. B. Du Bois Lectures*

\$12.50

## ***Tax Expenditures***

*Stanley S. Surrey and Paul R. McDaniel*

"This book is comprehensive, challenging, and constructive, and it fills a big void in the tax literature. It will interest economists, political scientists, lawyers, accountants, and legislators."—Joseph Pechman

\$27.50

## ***Housing and Neighborhood Dynamics***

*A Simulation Study*

*John F. Kain and William C. Apgar, Jr.*

Using the Harvard Urban Development Simulation Model, the authors assess the effects of programs that provide direct assistance to low-income property owners in an attempt to arrest neighborhood decline and encourage revitalization.

*Harvard Economic Studies*, 157

\$27.50



## ***Family Firm to Modern Multinational***

*Norton Company, a New England Enterprise*

*Charles W. Cheape*

The history of Norton Company, so carefully detailed here from company records and interviews, illustrates both the continuity and change important in the evolution of large-scale business in the United States.

*Harvard Studies in Business History*, 36

\$25.00

## ***Equality in America***

*The View from the Top*

*Sidney Verba and Gary Orren*

Based on a study of leaders from all significant sectors of American society, this book examines American attitudes toward equality by placing those beliefs in historical context and demonstrating a relationship between political and economic equality.

\$25.00 cloth; \$12.50 paper

## ***Tax Revolt***

*Something for Nothing in California*

*Enlarged Edition*

*David O. Sears and Jack Citrin*

"This book is the first substantial study of the California tax revolt. It is well crafted and well written and has great depth and sophistication. The topic is of intense interest and importance . . . A significant work."—Walter Dean Burnham

\$9.95 paper

## ***A Treatise on the Family***

*Gary S. Becker*

"This truly pathbreaking book marries techniques and problems hitherto regarded as utterly incompatible—rigorous economic reasoning to understanding the family. The marriage is astoundingly productive. It is destined to affect the foundations of every science dealing with human behavior."—Milton Friedman

\$7.95 paper



# **Harvard University Press**

79 Garden Street, Cambridge, Mass. 02138

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

# New From St. Martin's Scholarly

## **The Economics of Oil Crisis**

Theories of Oil Crisis, Oil Rent and Internationalization of  
Capital in the Oil Industry

Cyrus Bina

In this in-depth study of the causes and consequences of the 1972 crisis, the author, a former financial analyst with the treasury department of Iran, rejects the traditional theory that the large price jumps resulted from voluntary actions taken by OPEC and other Third World nations. He argues instead that it was actually the economics of domestic oil production in the continental United States that generated the huge price spiral.

1985

268 pp.

ISBN 0-312-23661-1

\$29.95

## **The Economic Theory of Multinational Enterprise**

Selected Papers

*Peter J. Buckley and Mark Casson*

Basing their approach on the internalization rubric and least cost location, the authors compare the economics theory of the multinational enterprise with alternative theories, presenting an authoritative review of the evidence that emerges, discuss alternatives to the multinational, including multiplant firms and cartels, examine the aspects of MNE behavior found to be common to all modes of international technology transfer, present a detailed modelling of the location decisions of MNEs, discuss the influence of financial and organizational factors, and integrate the concept of entrepreneurship into the theory, stressing the importance of flows of information and intermediate products.

1985

235 pp.

ISBN 0-312-23636-0

\$27.50

## **Free Market Economics**

A Critical Appraisal

*Andrew Schotter*

Schotter examines the theoretical bases of competitive economics, characterizing the free market economic argument, its basic assumptions and their roots in economic thought. He further presents criticisms of this argument, raising doubts about the blind advocacy of free market solutions to current social problems, and applies these criticisms to an analysis of public policy issues including minimum wage, affirmative action, crime and welfare, and the rational expectations debate. **"A clear, logical and penetrating analysis."**—*Choice*

1984

147 pp.

ISBN 0-312-30369-6

\$22.50

## **Full Employment Without Inflation**

Manifesto for a Governed Economy

*Tim Hazledine*

In this provocative study, the author develops an original reformulation of government's role in policy which, he contends, will lead to a stable and sensible economic environment. The issues discussed include the failure of Keynesianism and Monetarism, principles of elasticity and policy instruments, price controls, trade and exchange rates, deficits, monopoly, the labor force, and industrial, energy and environmental policy.

1984

264 pp.

ISBN 0-312-30971-6

\$27.50

## **Macroeconomics and Monopoly Capitalism**

*Ben Fine and Andy Murfin*

This book is a critique of current economic theory—Monetarism, Keynesianism and post-Keynesianism. The authors describe the economic characteristics of modern capitalism, emphasizing the role of giant corporations and governments, examine Keynesianism and Monetarism both in terms of their internal logic and their relevance to the past and future growth of the capitalist economy, and explore the strengths and weaknesses of the radical tradition associated with Baran, Sweezy, Kalecki and the post-Keynesians, discussing issues of monopolization and crisis. **"This book is a well-written critical appraisal. . . . Economists working within those traditions [discussed] should feel obligated to respond to the important arguments developed in this book."**—*Malcolm C. Sawyer,*

*University of York*

1984

170 pp.

ISBN 0-312-50336-9

\$22.50

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

## Markets, Entrepreneurs and Liberty

An Austrian View of Capitalism

W. Duncan Reekie

This book is a comprehensive survey of the main tenets and methodological approach of Austrian economics, exploring the uniqueness of Austrianism in economic thought and its relevance to economics as a whole in the late twentieth century. The author traces the evolution and growth of Austrian economics, and discusses such topics as the role of exchange, entrepreneurship and the theory of trade cycle, which offers one of the few satisfactory explanations of the high level of unemployment in today's economy.

1985

171 pp.

ISBN 0-312-51533-2

\$27.50

## Public Enterprise Economics

Second Edition

Ray Rees

This book is a fully updated and comprehensive analysis of the pricing, output and investment problems of public enterprises. Particular emphasis is placed on the actual dilemmas these institutions currently face and on the principles which would allow them to be run in a decentralized way. Topics discussed include the purposes and performance of public enterprises, welfare economics, marginal-cost pricing, taxation and income distribution. Appendices provide explanations of mathematical derivations used in the text.

1984

238 pp.

ISBN 0-312-65439-1

\$27.50

## Demand, Equilibrium and Trade

Essays in the Honor of Ivor F. Pearce

edited by A. Ingham and A. M. Ulph

This volume contains thirteen essays by leading economists that analyze Pearce's work on the theories of demand, general equilibrium, trade and capital, and economics as a whole; and offer critical reassessments and new findings in such areas as the Keynes-Ramsey Rule, Le Chatelier Principle, measures of complementarity and transfer pricing.

1984

256 pp.

ISBN 0-312-19187-1

\$37.50

## Expectations

Theory and Evidence

K. Holden, D. A. Peel and J. L. Thompson

This book is a synthesis of recent research and thought on the subject of expectations. The authors provide introductions to the use survey measures of expectations, the adaptive expectations approach, the concept of rational expectations, and the use of time series modelling. With this foundation, they discuss such topics as the implications of expectations in single-equation studies, the efficient markets hypothesis, the role of expectations in the consumption function and wage equation, the representation of expectations in both large- and small-scale macroeconomic models, the formation of aggregate expectations, and the combination and impact of forecasts on asset markets.

June

200 pp. (est)

ISBN 0-312-27599-4

\$25.00

## New in Paperback!

### Rational Expectations

An Elementary Exposition

G. K. Shaw

A basic introduction to the subject, written at a non-technical level. Shaw examines the need for expectations theory, and the static, adaptive and Keynesian treatment of expectations. He then considers rational expectations as a reaction to earlier work, its policy implications and its relation to Monetarism, business cycle theory and supply side economics. **"I am impressed by this as an introductory book for students. Professor Shaw succeeds in explaining rational expectations in a way that will be readily understood by undergraduate students."**—Professor J. Black, University of Exeter **"This is the only beginner's-level book on this subject of which this reviewer is aware."**—Choice

1984

131 pp.

ISBN 0-312-66403-6

\$9.95

## St. Martin's Press

Scholarly and Reference Books

175 Fifth Avenue • New York, NY 10010

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

The American Economics Association (AEA) is now soliciting applications to host the AEA Summer Program for Minority Students, for three summers beginning in 1986.

This program is now in its twelfth year and is currently at the University of Wisconsin at Madison. Previous host institutions have been Berkeley, Northwestern, and Yale.

The intent of the program is to increase the number of Blacks, Hispanics, and Native Americans pursuing the Ph.D. in Economics. In recent years the course of study has been an intensive eight to ten week program in intermediate microeconomics and macroeconomics, at the honor's level, and courses in econometrics or mathematical economics.

Funding for the program has been provided by the hosting institutions and grants to the AEA. Applications should be sent to Professor Donald J. Brown, chairman of the AEA Committee on the Status of Minority Groups in the Economics Profession, CSMGEP, no later than September 1, 1985.

CSMGEP  
Attention: Donald J. Brown  
Chairman, Department of Economics  
Yale University  
Box 1972 Yale Station  
New Haven, CT 06520-1972

# New from Rowman & Allanheld...

## **A Citizen's Guide to the New Tax Reforms**

### **Fair Tax, Flat Tax, Simple Tax**

**Edited by Joseph A. Pechman.** Gives average taxpayers the information they need to influence the national debate on the future of our tax laws.—**Senator Bill Bradley**

"An excellent guide to the major tax reforms now being considered by Congress."—**Congressman Jack Kemp**

"A superbly readable and reliable road map through the maze of tax reform."—**Walter Heller**  
176 pp. / ca \$21.50 / ca \$9.95 paper

## **The Economics of the Industrial Revolution**

**Edited by Joel Mokyr.** Having access to new economic and historical data, the writers challenge many popular interpretations of the effect of the Industrial Revolution and, in particular, offer new viewpoints on the standard-of-living controversy, the role of agriculture, the role of demand, and the sources of labor supply. Of special value is the in-depth introductory chapter by the editor.  
ca 352 pp. / ca \$34.50 / ca \$15.95 paper

## **Ethics, Efficiency, and the Market**

**Allen Buchanan.** An ideal critical synthesis of the best thinking on the role of the market as a basic institution of social organization. It articulates efficiency arguments and ethical arguments—and examines their conceptual, empirical, and moral presuppositions, as well as their implications for capitalist, socialist, and market-socialist economic arrangements.

"Buchanan offers a welcome dose of common sense and careful reasoning . . . a cogent demolition of several strong positions that are often asserted."

—**Russell Hardin**, The University of Chicago  
ca 180 pp. / ca \$24.50 / ca \$10.95 paper

## **The Demand for Primary Health Services in the Third World**

**John S. Akin et al.** A survey and analysis of the important issues connected with the availability (and the actual use) of health care in the Third World. The authors review the existing literature and apply sophisticated microeconomic theory and econometric methods to interpret new survey data.  
272 pp. / \$39.50

## COMING SOON...

### **Economic Justice**

#### **Private Rights and Public Responsibilities**

**Edited by Kenneth Kipnis and Diana T. Meyers**

A stimulating investigation into the nature of the economically just society. These papers set out the problems of contemporary social theory within the context of the distributive justice vs. property rights debate initiated by the works of John Rawls and Robert Nozick.

ca 325 pp. / ca \$28.50 / ca \$13.50 paper

### **The Impact of Inflation on Financial Activity in the U.S. Farming Sector**

**Y. Goldschmidt et al.** Major topics include interest charges on working capital, effect of debt finance on liquidity, impact of inflation of tax liability resulting from interest on loans, income measurement with a special emphasis on performance evaluation. Each of these discussions is followed by a thorough examination of applications to the US farm sector for the period 1950-1982, and the final chapter considers the important policy implications of this new research.  
ca 175 pp. / ca \$35.00

## NOW AVAILABLE IN PAPERBACK

### **Conversations with Economists**

#### **New Classical Economists and Their Opponents Speak Out on the Current Controversy in Macroeconomics**

**Arjo Klammer**

"Klammer's interviews are broad, incisive, and revealing. With great skill, he coaxes personal and economic insights out of his subjects . . . This is economic writing at its best—readable and human."—**Business Week**

"Brilliant."—**Leonard Silk**, The New York Times

**Interviewees:** Robert Lucas, Thomas Sargent, Robert Townsend, James Tobin, Robert Solow, Franco Modigliani, Alan Blinder, John Taylor, Karl Brunner, David Gordon, Leonard Rapping.

278 pp. / \$18.95 / \$9.95 paper

Write for our new 1985 Economics Catalog



**ROWMAN & ALLANHELD**

81 Adams Drive, Totowa, NJ 07512

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

## THIRD WORLD DEFAULT!

Two new books on the international debt situation focus on the dangers of default to the global economy.

**The Costs of Default**, by Anatole Kaletsky of London's *Financial Times*, argues that the debt burden on Third World countries has so limited their economic growth that political pressures for some form of default are building up. He suggests there may be a series of "conciliatory" defaults and calls for new negotiations leading to cooperation between debtor and creditor nations.

**The Debt of Nations**, by M. S. Mendelsohn, a financial writer who is a consultant to the World Bank and the Group of Thirty, argues that all involved have a stake in an orderly resolution of the crisis. He predicts that the debtor countries will manage to avert default.

These books cost \$7.00 each, plus \$1.00 postage and handling.

Please send prepaid orders to:

The Twentieth Century Fund • 41 East 70th Street • New York, NY 10021

New from the IMF . . .

### **Foreign Private Investment in Developing Countries**

#### ***A Study by the Research Department***

Since the early 1970s, foreign private investment in developing countries has been dwarfed as a source of capital by debt-creating bank credit. This shift may have rendered recipients more vulnerable to external difficulties; meanwhile, it is predicted that bank lending to these countries will grow relatively slowly over the medium term. Both these factors imply that foreign private investment could become more important.

Number 33 in the Occasional Paper Series of the International Monetary Fund, this paper examines the reasons for the decline of foreign direct and portfolio equity investment, its role in development and external adjustment, and, in this context, the policies of both recipient and source countries that might encourage larger flows.

Price: US\$7.50 (US\$4.50 to university libraries, faculty members, and students)

Available from: Publications Unit, Box E-197  
International Monetary Fund  
700 19th Street, N.W.  
Washington, D.C. 20431, U.S.A  
Telephone: (202) 473-7430

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*



# SUPPLEMENT YOUR MICRO, TRADE, AND PUBLIC FINANCE COURSES WITH GEMODEL

## A GENERAL EQUILIBRIUM MODEL

for the IBM PC, Apple II+ and TRS-80

GEMODEL is easy-to-use, menu-driven micro-computer software that solves real general equilibrium models in their non-linear form and:

- runs either as a 3-factor, 3-industry model or as a 2-factor, 2-industry model, with several consumer classes in both cases.
- solves within seconds closed and open economy models with or without international capital flows.
- illustrates at the touch of a key almost any aspect of Price, Trade, Taxation & Distribution theories. It is an indispensable tool for analysis of second-best problems. It will enrich your courses with discussion of numerical results.

GEMODEL uses nested CES utility and production functions to simulate the effects of changes in:

- factor endowments
- tariffs
- technology and productivity
- terms of trade
- tastes for goods and leisure
- interindustry relations
- and taxes.

GEMODEL simulates the allocative and distributive effects of a large array of taxes:

- corporate income tax
- social security and payroll taxes
- retail sales taxes
- progressive personal income tax
- property taxes
- taxes on sales of intermediate inputs
- and import duties;

as well as a similar array of subsidies and lump-sum redistributions of tax revenue.

GEMODEL will save your data and reload them for new simulation runs. Default values are provided for use without data input.

## We Provide User Support

For sample data, output, and further information, send \$5.00 (\$6.50 Cdn.) to the address below.

PRICE: \$395.00 U.S./\$495.00 CDN.

Residents of Ontario please add 7% provincial sales tax.

Please send me **GEMODEL**

Payment is enclosed by ( ) cheque ( ) money order

Name .....

Address .....

No. Street

City

State

ZIP

I use ( ) Apple ( ) IBM PC ( ) TRS-80

To order please mail this coupon. Make cheque or money order payable to:

**DAMUS INVESTMENT & AGENCY INC.**

1879 Kingsdale Ave., Ottawa, Ontario, K1T 1H9, CANADA

**DM**

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

# Economic Growth in the Third World, 1850-1980

Lloyd G. Reynolds

This book is the first to draw together into a systematic framework the increasing body of literature on the economic histories of individual third world countries. Focusing on the forty-one largest countries in Asia, Africa, and Latin America, Reynolds shows that the third world has a rich historical record of growth and that its growth patterns bear some resemblance to those observed earlier in Europe and North America. A Publication of the Economic Growth Center \$35.00

# What Is Political Economy?

*A Study of Social Theory and Underdevelopment*  
Martin Staniland

In "the most comprehensive and sensible guide to the political economy literature available" (William J. Foltz), Staniland explores current meanings of the term political economy, particularly as it is applied to situations in the third world. He critically analyzes a broad variety of theories, including rational choice theory, dependency theory, politicism, interdependence theory, and neo-Marxism. \$18.50

# Yale

Yale University Press  
Dept. 781J 92A Yale Station New Haven, CT 06520

# Policy, Power, and Order

*The Persistence of Economic Problems in Capitalist States*  
Kerry Schott

Inflation, unemployment, and slow growth have been persistent problems in all Western economies and neither Keynesian nor neoclassical policies seem able to remedy this. In this lucid book, Schott argues that the assumptions on which economic policy is based are naive and inconsistent. Criticizing both Marxist and individualist analyses of the role of the state, she stresses the fundamental importance for economists of political and social conditions and changing power relationships, and uses this to examine both the different investment and growth performances of various nations in recent years and the dilemmas now confronting economic policy makers. \$20.00

# Wayward Capitalists

*Target of the Securities and Exchange Commission*  
Susan P. Shapiro

This pathbreaking study describes the types, strategies, and impact of securities fraud and documents how these crimes are detected and handled by the SEC. Based on an inside look at SEC records over a twenty-five-year period, the book concludes with policy recommendations to improve the agency's work as the government's watchdog over the stock market. \$26.00

# Defending White-Collar Crime

*A Portrait of Attorneys at Work*  
Kenneth Mann

The first detailed description of what lawyers specializing in white-collar crime defense work do to prevent their clients from being indicted or convicted, this book raises important ethical questions about the nature of the legal profession in the United States. \$25.00

# United Nations Periodicals

## CTC REPORTER (CENTRE ON TRANSNATIONAL CORPORATIONS)

Reports on matters related to transnational corporations in all governmental and non-governmental organizations. Issued three times per year, including special supplement, in English, French and Spanish.

*Annual rate \$20.00*

*Single issues \$9.00*

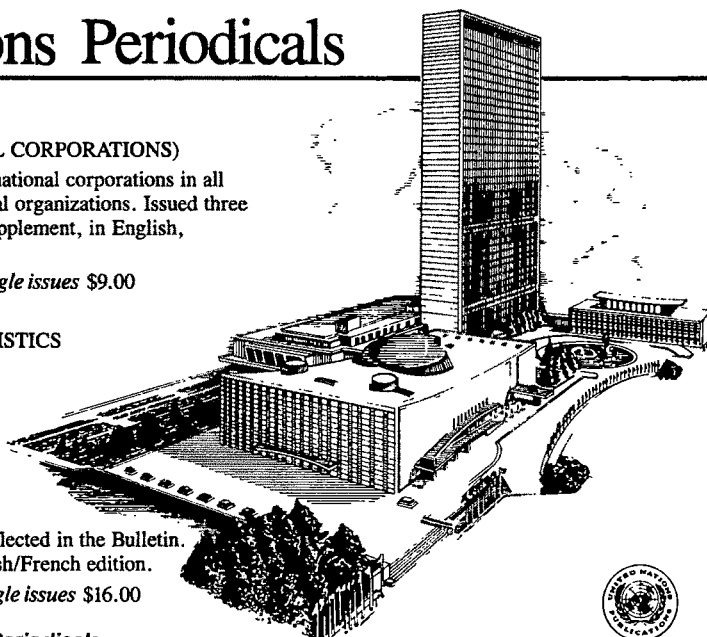
## MONTHLY BULLETIN OF STATISTICS

Provides monthly statistics on seventy-four subjects from over two hundred countries and territories, including special tables that graphically portray important economic developments. Quarterly data for significant world and regional aggregates are also reflected in the Bulletin. Issued monthly in a bilingual English/French edition.

*Annual rate \$150.00*

*Single issues \$16.00*

**For further information on all Periodicals of the UN send for our free Periodicals Brochure and our Catalogue of International Publications**



**UNITED NATIONS PUBLICATIONS**  
Room DC2-853, New York, N.Y. 10017  
Palais des Nations, 1211 Geneva 10, Switzerland

## THE ECONOMICS INSTITUTE

Boulder, Colorado  
offering

**Intensive Preparation For  
Post-Graduate Programs  
In Economics And Business  
For Students From Abroad**

- ☐ English      ☐ Mathematics      ☐ Economic Theory
- ☐ Statistics      ☐ Accounting      ☐ Computer Applications
- ☐ Business Organization and Management

*Current program dates, tuition and  
fee schedule, and further  
information available from:*

**The Director  
Economics Institute  
Campus Box 259  
University of Colorado  
Boulder, Colorado 80309**



Sponsored by the American Economic Association

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

## Computer Access to Articles in the JEL Subject Index

On-line computer access to the *JEL* and *Index of Economic Articles* data base of journal articles is currently available through DIALOG Information Retrieval Service. DIALOG file 139 (*Economic Literature Index*) contains complete bibliographic citations to articles from some 260 journals listed in the quarterly *JEL* issues from 1969 through the current issue. The abstracts in *JEL* since June 1984 are also available as part of the full bibliographic record. The file may be searched using free-text searching techniques or author, journal, title, geographic area, date, and other descriptors, including descriptor codes based on the *Index's* four-digit subject classification numbers.

DIALOG offers a variety of contract choices, including the option to pay for only what you use—*no minimum, no initiation or start-up fee*. Most university libraries already subscribe to DIALOG. For information on the DIALOG service, contact your librarian or write to or call:

DIALOG Information Services, Inc., Marketing Department, 3460 Hillview Avenue, Palo Alto, California 94304 (800/227-1927 or 800/982-5838, in California, or 415/858-3785).

# BUY 3

# GET 1 FREE!

## MONITORING GROWTH CYCLES IN MARKET-ORIENTED COUNTRIES

**Philip A. Klein and Geoffrey H. Moore**

*A National Bureau of Economic Research book*

Shows how the system of indicators developed over the past fifty years can be successfully applied to growth cycles, as well as to classical business cycles; shows further how this system can be applied to growth cycles in other countries with case studies of: Canada, the United Kingdom, the Federal Republic of Germany, France, Italy, Sweden, Belgium, the Netherlands, and Japan.

May 1985 — 288 pages

## THE LEVEL AND COMPOSITION OF HOUSEHOLD SAVING

**Patric H. Hendershott, Editor**

*Sponsored by the American Council of Life Insurance*

Analyzes the quantity and form of household saving, including wealth and portfolio changes of households, and specific economic decisions: purchases of consumer durable goods, the demand for pension savings, and investment in human capital such as education. Addresses the importance of corporate and government saving in the determination of household saving.

1985 — 328 pages

## AMERICAN JOBS AND THE CHANGING INDUSTRIAL BASE

**Eileen L. Collins and Lucretia Dewey Tanner, Editors**

Addresses the American worker's plight in the face of technological change. Details new patterns of trade, productivity, and industrial realignment in an effort to assess the future structure of U.S. employment.

1984 — 258 pages

## EXCHANGE RATES, TRADE, AND THE U.S. ECONOMY

**Sven W. Arndt, Richard J. Sweeney, and Thomas D. Willett, Editors**

*An American Enterprise Institute book*

Analyzes the data now available on the transition of the world's economy from fixed to floating exchange rates, considers the effects of flexible rates on international trade, and covers the macroeconomic linkages and influence of international affairs on the U.S. macro economy.

1985 — 328 pages

## REVITALIZING AMERICAN INDUSTRY

**Milton Hochmuth and William Davidson, Editors**

Argues for a national industrial policy to replace *ad hoc* and fragmented measures that can't protect us from tough newly industrialized competitors like Korea, Taiwan, Brazil, and Mexico.

1985 — 420 pages

## FREE WHEN YOU ORDER 3 TEN YEARS OF MULTINATIONAL BUSINESS

**Malcolm Crawford and James Poole, Editors**

*An Abt Book/ECONOMIST INTELLIGENCE UNIT special series*

Case studies of India, off-shore tax planning, currency swaps, SKF, Brown Boveri, McKinsey, DuPont, and major oil companies are included.

1982 — 184 pages

☐ **YES!** Please send me my **FREE** copy of **TEN YEARS OF MULTINATIONAL BUSINESS** (6609317) with the 3 books I've ordered below:

☐ No, I have not ordered a minimum of three books, but please send me:

- **AMERICAN JOBS AND THE CHANGING INDUSTRIAL BASE** (6609804) @ \$32.00 \$ \_\_\_\_\_
- **EXCHANGE RATES, TRADE, AND THE U.S. ECONOMY** (6609614) @ \$38.00 \$ \_\_\_\_\_
- **LEVEL AND COMPOSITION OF HOUSEHOLD SAVING** (6609242) @ \$32.00 \$ \_\_\_\_\_
- **MONITORING GROWTH CYCLES IN MARKET-ORIENTED COUNTRIES** (6610067) @ \$35.00 \$ \_\_\_\_\_
- **REVITALIZING AMERICAN INDUSTRY** (6609846) @ \$39.95 \$ \_\_\_\_\_

My state's sales tax \$ \_\_\_\_\_  
Postage/handling (\$1.50/bk) on charge orders \$ \_\_\_\_\_  
Prepaid orders are postage free!  
TOTAL \$ \_\_\_\_\_

☐ payment enclosed ☐ bill me  
charge my ☐ MC ☐ VISA ☐ AMX

Card No. \_\_\_\_\_ Exp. date \_\_\_\_\_

Signature \_\_\_\_\_

Send to: \_\_\_\_\_

ZIP \_\_\_\_\_

Prices subject to change.  
All orders subject to credit approval. U.S. funds only.  
Special offer valid to 30 June 1985. Special offer not available outside the U.S.A. and Canada. If you order by phone, tell the operator your order code is **AAER585**

**BALLINGER**  
PUBLISHING COMPANY  
Order Department  
2350 Virginia Avenue, Hagerstown, MD 21740  
To order, call toll free 1-800-638-3030

## MATHEMATICAL AND STATISTICAL PROGRAMMING PACKAGE FOR YOUR IBM PC

FAST • EASY TO USE • POWERFUL

# GAUSS™

YOU'VE NEVER SEEN ANYTHING LIKE IT!

**GAUSS** is a sophisticated mathematical and statistical programming package for the IBM PC and compatibles. It combines **speed, power, and ease of use** in one amazing program.

**GAUSS** allows you to do essentially anything you can do with a mainframe statistical package — and a lot more.

Personal computers are friendly, convenient, and inexpensive. So is **GAUSS**. **GAUSS** is not just a stripped-down mainframe program. **GAUSS** has been designed from the ground up to take advantage of all of the conveniences of a personal computer. After trying **GAUSS**, you may never use a mainframe again.

**GAUSS** comes with programs written in its matrix programming language that allow you to do most econometric procedures, including OLS, 2SLS, 3SLS, PROBIT, LOGIT, MAXIMUM LIKELIHOOD, and NON-LINEAR LEAST SQUARES.

In the current version, **GAUSS** will accept up to 90 variables in a regression. There is no limit on the number of observations.

**GAUSS** will do a regression with 10 independent variables and 800 observations in under 4 seconds — and with 50 variables and 10,000 observations in under 18 minutes. It will compute the maximum likelihood estimates of a binary logit model, with 10 variables and 1,000 observations, in 1-2 minutes, depending upon the number of iterations required.

**GAUSS** allows you to do complicated statistical procedures that you would never imagine trying on a mainframe. It is easy to program almost any routine, and **GAUSS** is so fast that it can do almost any job. But the nicest thing of all is that the cost of time on your personal computer is essentially zero!

**GAUSS** is an excellent teaching tool. It makes programming easy and allows students to focus on concepts and techniques.

If you can write it mathematically, you can write it in **GAUSS**. Furthermore, you can write it in **GAUSS** almost exactly the way you would write it mathematically.

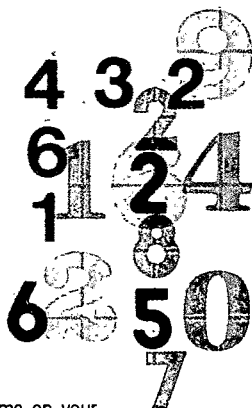
**GAUSS** is 10-15 times faster than other programs that use the 8087, and 15-100 times faster than other programs that do not use the 8087.

As in APL, single statements in **GAUSS** can accomplish what might take dozens of lines in another language. However, **GAUSS** provides you with additional powerful numerical operators and functions — especially for statistics and the solution of linear equations — that are not found in APL. And, of course, the syntax in **GAUSS** is much more natural (for most of us) than that in APL.

**GAUSS** has state-of-the-art numerical routines and random number generators.

**GAUSS** is extremely accurate. It allows you to do an entire regression in 19 digit accuracy. It will compute the Longley benchmark coefficients in 5 hundredths of a second with an average of 11 correct digits! (Try that on a mainframe!)

**GAUSS**, with its built-in random number generators and powerful functions and operators, is an excellent tool for doing simulations.



## GAUSS and the 8087 NUMERIC DATA PROCESSOR GIVE YOU MINICOMPUTER PERFORMANCE ON YOUR DESKTOP.

### SPECIAL INTRODUCTORY OFFER

With 30 Day Money

Back Guarantee ..... Reg. 395.00 **\$250.00**

**GAUSS** requires an IBM PC with at least 256K RAM, an 8087 NDP, 1 DS/DD disk drive, DOS 2.0 (or above).

IBM is trademark of IBM Corporation

Call or Write

**APPLIED  
TECHNICAL  
SYSTEMS**

P.O. Box 6487, Kent, WA 98064  
(206) 631-6679

# OECD publications

ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT

## **International Banks and Financial Markets in Developing Countries**

*Dimitri Germidis and Charles-Albert Michalet*

Examines whether international banks which establish themselves in developing countries contribute to the development of the host countries, and how they effect local banking systems. Also examines the impact of international offshore financial centers on development.

41 84 04 1

94 pages

ISBN 92-64-12635-X

\$13.00

## **Investing in Free Export Processing Zones**

*Antoine Basil and Dimitri Germidis*

Analyzes the role of free export processing zones in attracting export-oriented foreign investment, assesses their impact on the host country's economy, and explores the problem of the relations between foreign firms operating in the free zone and domestic industries.

41 84 05 1

68 pages

ISBN 92-64-12634-1

\$11.00

## **Main Economic Indicators: Historical Statistics 1964-1983**

The new edition of this information-packed volume presents 20-year series of data for such economic variables as GNP; production, deliveries, and stocks by industry; rate of capacity utilization; housing starts; retail and wholesale sales; employment and unemployment; producer and consumer prices; imports and exports, and much more. Data is provided for every OECD country. The base year used on indicators is 1980, and data is shown on a monthly, quarterly, and annual basis.

31 84 21 3

656 pages

ISBN 92-64-02606-9

\$35.00

## **National Accounts, 1984 Edition Volume II: Detailed Tables 1970-1982**

This newly revised and expanded annual includes tables for each OECD country showing main aggregates, private final consumption expenditure by type and purpose, gross fixed capital formation by kind of activity of owner, gross accounts for central government, accounts for state or provincial government, accounts for local government, accounts for social security funds, accounts for financial and non-financial corporate and quasi-corporate enterprises, accounts for households and private unincorporated enterprises, external transactions/current and capital accumulation accounts, capital finance accounts by sector, gross domestic product by kind of activity, cost components of value added, and profit shares and rates of return on capital. Values are expressed in national currencies.

30 84 03 3

555 pages

ISBN 92-64-02662-2

\$59.00

Available from:

### **OECD Publications and Information Center**

1750-E Pennsylvania Avenue, N.W.

Washington, D.C. 20006-4582 Tel.: (202) 724-1857

OECD



OCDE



# INDEX OF ECONOMIC ARTICLES

prepared under the auspices of  
*The Journal of Economic Literature*  
of the  
*American Economic Association*

- ✓ Each volume in the **Index** lists articles in major economic journals and in collective volumes published during a specific year.
- ✓ Most of the **Index's** volumes also include articles of testimony from selected congressional hearings in government documents published during the year.
- ✓ No other single reference source covers as many articles classified in economic categories as the **Index**.
- ✓ The 1977 volume contains over 10,500 entries.

## Currently available are:

Volume	Year Covered
XI	1969
XII	1970
XIII	1971
XIV	1972
XV	1973
XVI	1974
XVII	1975
XVIII	1976
XIX	1977
XX	1978
XXI	1979

*an  
indispensable  
tool for...*

**ECONOMISTS  
REFERENCE LIBRARIANS  
RESEARCHERS  
TEACHERS  
STUDENTS  
AUTHORS**

Future volumes will be published regularly  
to keep the series as current as possible.

**Price:** \$50.00 per volume (special 30% discount to  
AEA members)

Distributed by

**RICHARD D. IRWIN, INC.** Homewood, Illinois  
60430



# The American Economic Review

## ARTICLES

- M. MCALEER, A. R. PAGAN, AND P. A. VOLKER  
What Will Take the Con Out of Econometrics?
- E. E. LEAMER  
Sensitivity Analyses Would Help
- C. A. HOLT  
An Experimental Test of the Consistent-Conjectures Hypothesis
- J. HALTIWANGER AND M. WALDMAN  
Rational Expectations and the Limits of Rationality
- R. SCHMALENSEE  
Do Markets Differ Much?
- M. ESWARAN AND A. KOTWAL  
A Theory of Contractual Structure in Agriculture
- S. N. WIGGINS AND G. D. LIBECAP  
Oil Field Unitization
- R. C. FEENSTRA  
Anticipated Devaluations, Currency Flight, and Direct Trade Controls in a Monetary Economy
- J. AIZENMAN AND J. A. FRENKEL  
Optimal Wage Indexation, Foreign Exchange Intervention, and Monetary Policy
- M. L. KATZ AND C. SHAPIRO  
Network Externalities, Competition, and Compatibility
- J. S. FEINSTEIN, M. K. BLOCK, AND F. D. NOLD  
Asymmetric Information and Collusive Behavior in Auction Markets
- H. SIDER  
Unemployment Duration and Incidence: 1968-82
- A. GLAZER  
The Advantages of Being First
- W. E. COLE AND R. D. SANDERS  
Internal Migration and Urbanization in the Third World

SHORTER PAPERS: G. L. Salamon; S. Titman; E. Renshaw; W. Sander; R. G. Sheehan; D. Backus and J. Driffill; W. S. Comanor and H. E. Frech III; C. Dale and C. Gilroy; M. Dotsey; S. Ambler and R. McKinnon; H. N. Goldstein and S. E. Haynes; B. Falk and P. F. Orazem; B. Cornell; P. Honohan; J. Carmichael and P. W. Stebbing; R. Bookstaber and J. Langsam; R. W. Garrison; R. A. Heiner.

JUNE 1985



# THE AMERICAN ECONOMIC ASSOCIATION

●Printed at Banta Company, Menasha, Wisconsin. The publication number is ISSN 0002-8282.

●*THE AMERICAN ECONOMIC REVIEW* including four quarterly numbers, the *Proceedings* of the annual meetings, the *Directory*, and *Supplements*, is published by the American Economic Association and is sent to all members six times a year: March; May; June; September; semi-monthly, December.

**Regular member dues** for 1985, which include a subscription to both the *American Economic Review* and the *Journal of Economic Literature* are as follows:

\$35.00 if annual income is \$30,000 or less;

\$42.00 if annual income is above \$30,000, but no more than \$40,000;

\$49.00 if annual income is above \$40,000.

**Nonmember subscriptions** will be accepted only for both journals: Institutions (libraries, firms, etc.), \$100 a year; individuals, \$65.00. Single copies of either journal may be purchased from the Secretary's office, Nashville, Tennessee.

In countries other than the United States, add \$11.00 to the annual rates above to cover extra postage.

●Correspondence relating to the *Directory*, advertising, permission to quote, business matters, subscriptions, membership and changes of address should be sent to the Secretary, C. Elton Hinshaw, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786. Change of address must reach the Secretary at least six (6) weeks prior to the month of publication. The Association's publications are mailed second class.

●Second-class postage paid at Nashville, Tennessee and at additional mailing offices. Printed in U.S.A.

●Postmaster: Send address changes to *American Economic Review*, 1313 21st Avenue So., Suite 809, Nashville, TN 37212-2786.

Founded in 1885

## Officers

### *President*

CHARLES P. KINDLEBERGER

Massachusetts Institute of Technology

### *President-Elect*

ALICE M. RIVLIN

The Brookings Institution

### *Vice Presidents*

ELIZABETH E. BAILEY

Carnegie-Mellon University

JOSEPH E. STIGLITZ

Princeton University

### *Secretary*

C. ELTON HINSHAW

Vanderbilt University

### *Treasurer*

RENDIGS FELS

Vanderbilt University

### *Managing Editor of The American Economic Review*

ORLEY ASHENFELTER

Princeton University

### *Managing Editor of The Journal of Economic Literature*

MOSES ABRAMOVITZ

Stanford University

## Executive Committee

### *Elected Members of the Executive Committee*

WILLIAM D. NORDHAUS

Yale University

A. MICHAEL SPENCE

Harvard University

VICTOR R. FUCHS

Stanford University

JANET L. NORWOOD

Bureau of Labor Statistics

ALAN S. BLINDER

Princeton University

DANIEL L. McFADDEN

Massachusetts Institute of Technology

### *EX OFFICIO Members*

W. ARTHUR LEWIS

Princeton University

CHARLES L. SCHULTZE

The Brookings Institution

## EVSEY D. DOMAR

DISTINGUISHED FELLOW

1984

In a celebrated series of papers, beginning in 1946 and extending over a decade, Evsey Domar became one of the founding fathers of the modern study of economic growth. He and Roy Harrod—whose point of view was rather different, although the adjective “Harrod-Domar” entered our lexicon for good—revived for our time the grand long-term dynamical questions that preoccupied the great economists of the eighteenth and nineteenth centuries. What is more, they insisted that the investment requirements of long-term growth were an essential backdrop to and causal factor in shorter-run economic fluctuations.

Evsey Domar’s second career was in the study of comparative economic systems, as befits a scholar who arrived on the East Coast of the United States by way of Manchuria. As a student of the economy of the USSR as an analyst of the cooperative firm, and as the propounder of an extraordinarily fruitful historical generalization about the origin of slave and serf systems, his example has taught a generation of economists the value of a broad historical, geographical, and institutional imagination.



*Grey D. Roman*

# THE AMERICAN ECONOMIC REVIEW

June 1985

**Managing Editor**  
ORLEY ASHENFELTER

**Co-Editors**  
ROBERT H. HAVEMAN  
JOHN G. RILEY  
JOHN B. TAYLOR

**Production Editor**  
WILMA ST. JOHN

**Board of Editors**  
GEORGE A. AKERLOF  
CLIVE BULL  
MICHAEL R. DARBY  
JACOB A. FRENKEL  
CLAUDIA D. GOLDIN  
PHILIP E. GRAVES  
GEORGE E. JOHNSON  
JOHN F. KENNAN  
MERVYN A. KING  
MEIR KOHN  
PAUL KRUGMAN  
BENNETT T. McCALLUM  
EDGAR O. OLSEN  
RICHARD SCHMALENSEE  
STEVEN SHAPELL  
JOHN B. SHOVEN  
SUSAN WOODWARD  
LESLIE YOUNG

• Submit manuscripts (4 copies), no longer than 50 pages, double spaced, to:  
Orley Ashenfelter, Managing Editor, *AER*;  
169 Nassau Street, Princeton, NJ 08542-7067.

• Submission fee: \$25 for members; \$50 for nonmembers. Style guides will be provided upon request.

• No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

• Copyright © American Economic Association 1985. All rights reserved.

VOLUME 75, NUMBER 3

## Articles

- What Will Take the Con Out of Econometrics?  
*Michael McAleer, Adrian R. Pagan,  
and Paul A. Volker* 293
- Sensitivity Analyses Would Help  
*Edward E. Leamer* 308
- An Experimental Test of the Consistent-Conjectures Hypothesis  
*Charles A. Holt* 314
- Rational Expectations and the Limits of Rationality: An Analysis of Heterogeneity  
*John Haltiwanger and Michael Waldman* 326
- Do Markets Differ Much? *Richard Schmalensee* 341
- A Theory of Contractual Structure in Agriculture  
*Mukesh Eswaran and Ashok Kotwal* 352
- Oil Field Unitization: Contractual Failure in the Presence of Imperfect Information  
*Steven N. Wiggins and Gary D. Libecap* 368
- Anticipated Devaluations, Currency Flight, and Direct Trade Controls in a Monetary Economy  
*Robert C. Feenstra* 386
- Optimal Wage Indexation, Foreign Exchange Intervention, and Monetary Policy  
*Joshua Aizenman and Jacob A. Frenkel* 402
- Network Externalities, Competition, and Compatibility  
*Michael L. Katz and Carl Shapiro* 424
- Asymmetric Information and Collusive Behavior in Auction Markets  
*Jonathan S. Feinstein, Michael K. Block,  
and Frederick D. Nold* 441
- Unemployment Duration and Incidence: 1968-82  
*Hal Sider* 461
- The Advantages of Being First *Amihai Glazer* 473
- Internal Migration and Urbanization in the Third World  
*William E. Cole and Richard D. Sanders* 481

## Shorter Papers

Accounting Rates of Return	Gerald L. Salamon	495
Urban Land Prices under Uncertainty	Sheridan Titman	505
A Note on Equity and Efficiency in the Pricing of Local Telephone Services	Edward Renshaw	515
Women, Work, and Divorce	William Sander	519
Money, Anticipated Changes, and Policy Effectiveness	Richard G. Sheehan	524
Inflation and Reputation	David Backus and John Driffill	530
The Competitive Effects of Vertical Agreements?	William S. Comanor and H. E. Frech III	539
Enlistments in the All-Volunteer Force: Note	Charles Dale and Curtis Gilroy	547
Is There an Operational Interest Rate Rule?	Michael Dotsey	552
U.S. Monetary Policy and the Exchange Rate:		
Comment	Steven Ambler and Ronald McKinnon	557
Reply	Henry N. Goldstein and Stephen E. Haynes	560
The Money Supply Announcements Puzzle:		
Comment	Barry Falk and Peter F. Orazem	562
Reply	Bradford Cornell	565
Fisher's Paradox:		
Comment	Patrick Honohan	567
Reply	Jeffrey Carmichael and Peter W. Stebbing	569
Predictable Behavior:		
Comment	Richard Bookstaber and Joseph Langsam	571
Comment	Roger W. Garrison	576
Reply	Ronald A. Heiner	579
Auditors' Report		586
Notes		593

## Editor's Note

Manuscripts submitted for publication in the *American Economic Review* after March 1, 1985, will be handled by a new procedure. Each paper will be assigned to one of three Co-Editors, or retained by the Managing Editor, for supervision through the review process *and* for a final decision on the manuscript status. Robert Haveman, University of Wisconsin-Madison, John Riley, University of California-Los Angeles, and John Taylor, Stanford University, have agreed to join me as Co-Editors in this new endeavor.

It is my hope that our new editorial procedure will allow due recognition in the editorial process for the wide variety in both the substantive and methodological character of manuscripts submitted to the *Review*. It should also be possible to obtain these benefits without a significant increase in the length of time required to review and process submitted manuscripts.

I am especially pleased to report that Wilma St. John, Production Editor of the *Review* for seventeen years, has agreed to continue in that capacity at the new editorial office of the *Review* at 169 Nassau Street, Princeton, NJ 08542-7067.

Orley Ashenfelter  
Managing Editor

# What Will Take the Con Out of Econometrics?

By MICHAEL MCALEER, ADRIAN R. PAGAN, AND PAUL A. VOLKER\*

More than twenty years ago Carl Christ recounted a story about a new typist who rendered "econometrics" as "economic tricks." No doubt this tale was greeted with some amusement at that time; equally without doubt, today it would probably only occasion wry and knowing smiles. Charting the course of this transition, and accounting for its direction, has been the concern of a number of recent articles. Perhaps the most perceptive of these has been Edward Leamer's article (1983). His contribution is of special interest, as it seeks not only to be descriptive but prescriptive; methods are outlined, the use of which Leamer sees as essential to the restoration of confidence in econometric research. Such techniques have now been promulgated and applied in a number of contexts. Thomas Cooley (1982), for example, looks at the impact of industry concentration upon profits; Louis Dicks-Mireaux and Mervyn King (1984) consider the effect of pensions on savings; Cooley and Stephen LeRoy (1981) are concerned with money demand. These constitute just three of the more prominent applications.

Although the number of applications of the methods is growing, and approving references are being made to them, surprisingly few have queried the basis of the contention that the procedures really do allay some of the suspicion greeting econometric results; a singular exception being Phoebus Dhrymes

(1982). Yet the claims being made for this methodology are such as to demand a close investigation. As witnesses to these claims we quote Leamer and Herman Leonard:

We propose that researchers be given the task of identifying interesting families of alternative models and be expected to summarize the range of inferences which are implied by each of the families. When a range of inferences is small enough to be useful and when the corresponding family of models is broad enough to be believable, we may conclude that these data yield useful information. When the range of inferences is too wide to be useful, and when the corresponding family of models is so narrow that it cannot credibly be reduced, then we must conclude that inferences from these data are too fragile to be useful. ...

...The proper test of our proposals is whether they are useful in practice. We believe that researchers will find them to be efficient tools for discovering the information in data sets and for communicating findings to the consuming public. [1983, p. 306]

The aim of our paper is to consider possible answers to the question in the title. Because of its position as one proposed answer, and its strong advocacy by a number of authors (for example, Cooley and LeRoy, p. 827), we pay particular attention to Leamer's *Extreme Bounds Analysis (EBA)*. In our inquiry, contained in Sections I, II, and III, the discussion is structured along the lines of the three themes in the statement above: the effect of looking at different families of models, the determinants of a fragile inference, and the nature of the conclusions that may be drawn from the information provided by *EBA*. Based on the arguments of those sec-

\*McAleer and Pagan: Department of Statistics, The Faculties, Australian National University, Canberra, A.C.T. 2600, Australia; Volker: Bureau of Labour Market Research, Canberra. An earlier version of this paper appeared as a Working Paper in Economics and Econometrics, No. 097, Australian National University. It is available on request. We thank all those who commented on that version. We believe that those comments sharpened our arguments considerably. Special thanks go to Ed Leamer, Tom Cooley, Trevor Breusch, David Hendry, Allan Gregory, Hashem Pesaran, Peter Schmidt, and Pravin Trivedi.

tions, we conclude that *EBA* does not go very far in removing the con from econometrics. Furthermore, in most instances, it can actively distract a researcher from asking important questions about an econometric model.

But just because the promise and the performance of *EBA* diverge, it does not obviate the need for a methodology aiming to dispel doubts arising over conventional research presentation and analysis. Accordingly, Section IV sets out our own prescription, the basic ingredients of which are the necessity for a clear and full disclosure of the process whereby a preferred model was selected, and the requirement that a thorough evaluation has been made of the properties of such a specification. Such an orientation is scarcely original, reflecting in its concerns an oral tradition that owes much to Denis Sargan's (1964) influential paper on wages and prices. Nevertheless, it is worth explicitly stating these principles, as our experience convinces us that, consistently applied, they can go a long way towards the "de-conning" of econometrics. As an example of this approach, and to contrast our prescription with *EBA*, Section IV below re-examines the conclusions drawn by Cooley and LeRoy from their demand for money study.

### I. The Problems in Families

Trying to define "the family" nowadays is enough to give a sociologist a nervous breakdown. To keep things simple one is inclined to assign a few individuals to its core and then to generate a whole range of alternatives by adding on children, grandparents, aunts, uncles, and other "relatives." Such a homely analogy captures rather nicely the essence of the "family of models" mentioned by Leamer and Leonard. At their core are variables classified as *important*. Added on are variables termed *doubtful*. What demarcates them is that *only the latter can be combined in an arbitrary linear fashion*. We emphasize this definition, since much of the discussion and use of *EBA* tends to proceed as if the division were based upon whether the associated coefficients are likely to be

zero or not.<sup>1</sup> Because this is not so, the decision to assign variables to their respective classifications is not a trivial one, and we explore it in detail in Section III.

To complete the elements of *EBA*, the concept of a *focus* variable is needed. This derives from the assumption that the magnitude of one of the model coefficients is of special interest. By itself, this feature tells us nothing about the nature of such a variable; it may be free or doubtful. Examples of both are given by Leamer (1983). His "bleeding heart liberal" regards the impact of execution probability upon murders as doubtful, while his "right winger" treats the same variable as free.

Proponents of *EBA* work with the maximum and minimum point estimates of the focus coefficient as the set of restrictions upon the doubtful variables is changed. If the gap between these values is wide, readers are generally informed that no reliable inference can be drawn about the influence of the focus variable. Thus Cooley and Le Roy express the belief that almost nothing can be said about the value of the interest elasticity of the demand for money. Within one of their families of models this elasticity could lie anywhere between  $-6.27$  to  $2.24$ .

Now it is a rare family that does not have a member with problems at some stage or other. Families of models also share this characteristic, but this is rarely mentioned by *EBA* advocates. Notwithstanding that, it has to be the case that a consumer of the conclusion drawn from an application of the methodology must take some notice of the nature of the model that generated the extreme bounds. When this is done, there are at least two situations in which inferences drawn from *EBA* would have to be heavily discounted.

<sup>1</sup>Unfortunately, the terminology of "important" and "doubtful" tends to bolster this impression. For this reason, we substituted "free" for important as that captures the nature of these variables much more closely. Ideally, a similar change would have been desirable for doubtful.



First of all, the restrictions that are being imposed upon the doubtful variables may be entirely unacceptable. Suppose that  $\gamma_1$  and  $\gamma_2$  are the parameters associated with the income and lagged dependent variable terms in a money demand function, and both variables are treated as doubtful. Then a restriction of the form  $\gamma_2 - \theta\gamma_1 = 0$ , with  $\theta$  negative, would offend against theoretical conceptions. An extreme bound generated with  $\theta < 0$  in a money demand example would be of little interest and, yet, there is nothing to safeguard against such a possibility. While Leamer himself is aware of this problem (see Leamer, 1978, p. 199), there have been few attempts at cautioning users of *EBA* about it. Cooley and LeRoy do not mention it at all, despite the fact that income and wealth elasticities associated with one of the extreme bounds of the ninety-day Treasury bill rate are actually negative.

Attempts have been made to limit such conflicts. Leamer (1982) restricts the feasible parameter space by requiring an investigator to put upper and lower limits on prior variances. It is hard to know what to make of this "solution," as the choice of such limits is extremely difficult and essentially arbitrary. One person's view of what constitutes a reasonable bound is unlikely to coincide with another's, and there is always the residual suspicion that prior variances have been chosen to yield narrow or tight bounds. As a satisfactory alternative to current practice it leaves much to be desired. It re-introduces the very element of whimsy that *EBA* was supposed to ameliorate.

A second alternative is to ensure that the extreme bounds do not disagree too greatly with the sample. Cooley does this by invoking the constraint that estimates should lie within the  $\alpha$  percent confidence ellipsoid associated with the least squares estimates of the complete model. A range of sample-modified bounds can then be generated by varying  $\alpha$ . When  $\alpha = 100$ , the ordinary extreme bounds are found.

Once we introduce the sample evidence to constrain the alternative models, we are implicitly being asked to accept a number of conventions underlying *EBA* (at least as pre-

sented in the literature). Namely, that the errors in models be normally distributed, nonautocorrelated, and homoscedastic; that the regressors be predetermined; and finally, that sample sizes are large enough for "confidence intervals" to be known with accuracy. No doubt these conventions may be appealing, but Leamer himself has pointed out the problem with their use: "Though the use of conventions does control the whimsy, it can do so at the cost of relevance" (1983, p. 38). That principle is certainly applicable here, as the breakdown of any of these conventions means that the " $\alpha$  percent confidence intervals" are anything but, and exactly what constraint is being applied becomes increasingly hazy.

These considerations emphasize the absolute necessity to know the point estimates of all coefficients in the model generating the bounds. But such knowledge is still not sufficient to decide if we have just come across a problem child or not. It is perfectly possible for all point estimates to appear reasonable, but for the model to be rejected on other grounds, such as when it exhibits substantial serial correlation. An extreme value generated from such a model would not be of great interest, since an investigator would not regard it as a suitable candidate for conveying information about the focus coefficient. Without knowing the *full set of characteristics* of the models generating the extremes, it is impossible to know what weight should be placed on the latter. Mere provision of the bounds, as in Cooley and LeRoy for example, is not enough. Much more information is needed to assess whether these bounds are meaningful.

## II. When is an Inference Fragile?

In what has transpired so far we have been somewhat vague about exactly how the bounds are used to conclude that an inference is fragile. If left that way, *EBA* becomes a "black box," and no understanding of the factors leading to an inference being fragile would be available. For this reason, we have gleaned two interpretations of fragility from the literature applying *EBA*, each of

which is sufficiently precise to enable analytical results to be established.

The first of these, henceforth referred to as Type *A* fragility, is given by Leamer and Leonard as follows:

These extreme values,  $\hat{\beta}_{\min}$  and  $\hat{\beta}_{\max}$ , delineate the ambiguity in the inferences about  $\beta$  induced by the ambiguity in choice of model. If the interval  $[\hat{\beta}_{\min}, \hat{\beta}_{\max}]$  is short in comparison to the sampling uncertainty, the ambiguity in the model may be considered irrelevant since all models lead to essentially the same inferences. [p. 307]

With the sampling uncertainty measured as  $k$  times the estimated standard deviation of the focus coefficient,  $k$  being a predetermined constant, such a definition has been adopted by Leamer-Leonard, Cooley, and Cooley-LeRoy. The first provide no guidance about  $k$ , Cooley selects a value of  $k=4$ , while the last opt for  $k=2$ . To investigate the consequences of adopting this definition of fragility, we provide Proposition 1 (proof available on request).

#### PROPOSITION 1:

(a) *When the focus variable is doubtful, the necessary and sufficient condition for Type A fragility to exist is that the chi-square statistic for the doubtful variable coefficients to equal their prior means ( $\chi_D^2$ ) exceeds  $k^2$ .*

(b) *When the focus variable is free, the necessary condition for Type A fragility is that  $\chi_D^2 > k^2$ .*

Proposition 1 is quite striking, as it shows that whether an inference is to be Type *A*-fragile or not depends upon two quantities: namely, the significance of the doubtful variables in the model and the value chosen for  $k$ . Regarding the first, its magnitude will depend crucially upon the prior means assumed for the doubtful variables. If the prior means are taken to be zero, whereas the *OLS* estimates lie a long way from that point, a large value of  $\chi_D^2$  is likely. The closer the means are to the *OLS* values, the smaller will be  $\chi_D^2$ , and the less the evidence of

fragility. Everything therefore depends upon the whimsy of the reporter in the choice of prior means for doubtful variables! Hardly a good method for getting rid of the con artists. Instead it gives them enormous scope for generating almost any result they like. In the examples of *EBA* usage available, only D. G. Fiebig (1981) attempts to spell out this sensitivity of bounds to prior mean specification.

Proposition 1 moreover tells us something else of importance: that inferences will only be fragile if doubtful variables are informative. Assuming for convenience that prior means are zero, a large value of  $\chi_D^2$  signals to a researcher that these variables should appear in any model from which inferences are to be drawn. From this perspective, *EBA* is just an inefficient (and incomplete) way of communicating to readers the fact that the doubtful variables are needed to explain the data; a better solution would be to just present estimates of the general model along with an associated  $\chi_D^2$  statistic, letting consumers of research judge whether any further simplification of the model is justified.<sup>2</sup>

The analytic results presented in Proposition 1 can also rationalize the findings of a number of different investigations in which *EBA* has been employed. Leamer and Leonard's nuclear reactor example treats as doubtful those variables with *t*-values all below 1.03, leading to a lack of Type *A* fragility. In contrast, Cooley's profits regressions exhibit four variables with *t*-statistics greater than 3.5, and three of the four *always* appear as doubtful variables. Is it any wonder then that he concludes that Type *A* fragility exists for a concentration/profits relationship?

Perhaps the ambiguities raised above could be dissipated by an alternative definition of

<sup>2</sup>It is of interest to specialize  $k$  to 2. When only a single doubtful variable is present Type *A* fragility occurs when the *t*-statistic of the doubtful variable exceeds 2, which is a conventional rule of thumb for selection of regressors. As the number of doubtful variables grows, however, a constant value of  $k=2$  means that the comparison of  $\chi^2$  with 4 corresponds to larger and larger levels of significance. Most researchers would presumably find this implicit assumption in *EBA* a little odd.

fragility. Leamer and Leonard provide just that, and we will designate it as Type *B* fragility in what follows. They say: "An alternative definition of shortness derives from a decision problem based on  $\hat{\beta}$ : the interval is short if all values in the interval lead to essentially the same decision" (p. 307, fn. 1).

When implemented by Leamer (1983), Type *B* fragility occurs if there is a sign change implicit in the bounds. Ignoring, as Leamer does, the fact that these bounds themselves have standard errors, we proceed to analyze the nature of this type of fragility using Proposition 2 (proof available on request).

#### PROPOSITION 2:

(a) *When the focus variable is doubtful the necessary and sufficient condition for Type B fragility to exist is that  $\chi_D^2 > \chi_{F0}^2$ , where  $\chi_{F0}^2$  is the  $\chi^2$  statistic for testing if the focus coefficient is zero.*

(b) *When the focus variable is free, the necessary condition for Type B fragility is  $\chi_D^2 > \chi_{F0}^2$ .*

The movement from Type *A* to Type *B* fragility only changes the benchmark against which the significance of the doubtful variables is checked. It is no longer set by the reporter but determined by the data ( $\chi_{F0}^2$ ). Furthermore, when the focus variable is doubtful, it is always the case that  $\chi_D^2$  exceeds  $\chi_{F0}^2$  (ignoring singularities in the design matrix), and so Type *B* fragility is in evidence. While such a result is solely a consequence of the fact that zero is an admissible value for that doubtful variable coefficient, it serves to emphasize how Type *B* fragility may evenuate purely by a classification of variables. An example of this is provided by Leamer in his discussion of the impact of execution on murders. After placing the execution variable in the doubtful class, thereby producing an opposite sign to that from unrestricted least squares, he concludes: "I come away...with the feeling that any inference from these data about the deterrent effect of capital punishment is too fragile to be believed" (1983, p. 42).

Since the sign change did not depend in any way upon the data, we find such a conclusion a trifle hard to defend.<sup>3</sup>

#### III. When is a Variable Doubtful?

Propositions 1 and 2 strongly suggest that the conclusions on fragility drawn from *EBA* are intimately bound up with the classification of variables as doubtful and free. The polar case where the focus variable is orthogonal to all other regressors gives a striking demonstration of that fact. When treated as free, the gap between the bounds is zero, as the point estimate of the focus coefficient is entirely insensitive to combinations of other variables. But, when treated as doubtful, the width of the bounds varies directly with  $\chi_D^2$ ; the more significant the focus variable the greater the degree of fragility inferred.

A concrete example may serve to highlight just how important this choice can be to the outcome. Accordingly we consider the model of murder rates set out in the April 1983 SEARCH manual (it resembles that in Leamer, 1982). Table 1 gives the extreme bounds, range (the absolute value of the difference between the bounds), and ratio of range to standard errors for the impact of execution on murders under different variable designations.

As the definition of Type *A* fragility reflected the relative magnitudes of the range and standard deviation, the last column of Table 1 contains the information that would be used to assess whether inferences about the impact of executions upon murders are fragile. Clearly the decision about which variables are doubtful can have enormous consequences for any conclusions. Such variation naturally poses the question of how we are to know which one of the four options is

<sup>3</sup>Note that a sign change also occurred when the execution variable was free under the "eye-for-eye" specification. With eleven doubtful variables, and a *t*-statistic of less than two for the execution coefficient, an application of Proposition 2(b) should leave us in little doubt over why that was so.

TABLE 1—EXTREME BOUND INFORMATION FOR EXECUTION COEFFICIENT (*PX*)

Free Variables	Min	Max	Range	Range/ <i>SD</i>
None	-2.87	2.72	5.59	115.0
<i>PX</i>	-.45	1.35	1.8	37.0
<i>PX</i> , intercept	-.40	.10	.5	10.3
<i>PX</i> , intercept, other variables with $t > 3(S, PC,$ <i>PCTPOOR</i> )	-.22	-.01	.21	4.3

Note: *PC* = probability of conviction, *PX* = probability of execution, *S* = sentence, *PCTPOOR* = percent poor, standard deviation of focus coefficient = .0486, *SD* = standard deviation.

to be adopted? Or, when is a doubtful variable doubtful? The answer must be that there is no answer. A decision to nominate a variable as doubtful is a personalized one, resting very much upon the opinions and values of the nominator. Consensus is no more likely over this choice than in the traditional selection of regressors problem.

Having elicited this point, the most serious defect in *EBA* becomes transparent: unless extreme bounds are presented for *all* possible classifications of variables as doubtful and free, an observer cannot be certain that the selection does not constitute a "con job." *Selectivity in regression reporting therefore has as an exact analog in EBA the different classifications of variables as doubtful and free.* *EBA* users report results for only particular variable categories and so are as arbitrary and selective in their modus operandi as the practices they criticize and claim to be improving on.

We can see this effect in Table 1. With nine variables in the regression there are 181,440 possible doubtful/free splits. Hence, inevitably some selection from this huge number will be made. Someone intent on demonstrating that executions deter murders would undoubtedly quote the final classification (or an augmented version), while those wishing to denigrate such a position would opt for the first two doubtful variable choices. There seems no reason to suppose that all of the classifications in Table 1 would be given by either protagonist, any more than one would anticipate each individual presenting the equivalent set of regressions composed of the different types of free variables. Thus

there is little reason to believe that *EBA* provides a reporting style that is any better than that currently practiced.

Sections I–III can now be drawn together to highlight the fact that *EBA* is not a satisfactory solution to the question posed in the title of this paper.<sup>4</sup> Section I argued that the extreme bounds themselves are not enough to enable conclusions to be drawn regarding fragility; we need to know the characteristics of the models generating such bounds. Sections II and III demonstrated that *EBA* is as capable of manipulation as the traditional presentation it aims to replace; perhaps more so in one respect in that an additional arbitrary choice of prior mean must be made. Consequently, if one feels unhappy with the information provided by *selective regressions*, one should not be any more satisfied with extreme bounds obtained by *selective variable partitions*. A con man in one mode would have no fear of becoming deskilled in the other.

#### IV. Cooley and LeRoy's Demand for Money Function: Contrasting the Methodologies

Given our belief that *EBA* cannot de-con econometrics, is there anything that might?

<sup>4</sup>It is important to emphasize that an answer to this question is our central concern. We do not quibble with the contention that *EBA* displays the impact of prior information on posterior means. To do so would be inconsistent with our Proposition 1. Nor do we argue that, for a given variable partition, *EBA* might not be useful. In Section IV we do, in fact, exploit it in exactly such a context.

Not generally, as there are almost certainly instances in econometrics, just as in science, of outright fraud. Nothing will detect such deception, except a vigorous critical tradition and a requirement that utilized data be either available or easily replicable. But our perception of the skepticism greeting many econometric studies is that it does not arise from a high incidence of such a phenomenon. Rather it stems from a feeling that the sins are venial rather than mortal; something has been left undone that should have been done.

Now *EBA* clearly addressed itself to this problem by indicating, for a given variable partition and universe of variables, the worst outcomes if everything conceivable were done. What it leaves undetermined is both the process by which the partition it is conditional upon was arrived at, and the operating characteristics of models generating the extremes. Three points therefore always need to be considered in assessing an *EBA*. In turn, these three elements also occur in the traditional line of research and are, we believe, the source of much of the dissatisfaction with it. Because they are pivotal to the methodology advanced in this section, we list them below:

- A. Selection of a general model;
- B. How and why any general model was simplified to the preferred one(s);
- C. Quality control of the preferred model(s).

Selection of a general model is a problem with all research methodologies (including *EBA*) and we can do no better than concur with Leamer and Leonard when they say: "But it is up to readers of research to decide if the reported family of models is credibly inclusive. If the researcher, for whatever reason, selects an incredibly narrow family of models, readers will properly ignore the results" (p. 307).

Even if we largely agree that the choice of variables considered in an investigation was commendably large, it is frequently the case that little discussion is provided of the strategy employed to obtain a more parsimonious representation of the data. Where a systematic reduction is possible, it should be followed; where it is not, detail should be

sufficient to enable a consumer of the research to determine exactly the criterion adopted in performing the simplification. At a very minimum this forces the presentation of an estimated general model and some analysis of how the preferred model relates to it.

Our final category focuses upon the quality control exercised on the models presented. Frequently, this is little short of abysmal and, as James Ramsey comments, "...it is amazing that so little is done to evaluate the model and the results" (1983, p. 242). Yet, ultimately, quality control is as important for the econometrics profession as it is for automobile manufacturers. A gradual realization of this point has in fact stimulated the development of criteria for the formal evaluation of models. For later reference it is useful to summarize the outcome of that research by classifying derived methods into five major categories:

- 1) Consistency with theory; 2) Significance, both statistical and economic; 3) Indexes of inadequacy; 4) Fragility or sensitivity; 5) Whether a model can encompass or reconcile previous research.

These five categories can be viewed as a regrouping of the criteria suggested in David Hendry and Jean-François Richard (1982) for settling upon a "tentatively adequate" model. Categories 1 and 2 have tended to dominate in past evaluative analysis and even now constitute the corpus of most applied econometrics courses and texts. Increasing attention has, however, been paid to the necessity of item 3, with Hendry (1980) giving a general perspective and Hendry (1983) a detailed application. Robert Engle (1982b) and Pagan and Anthony Hall (1983) provide an account of much of the technology, emphasizing that these methods aim to extend the horizon in directions where errors might be anticipated. Some indexes, such as the Durbin-Watson statistic, have been routinely used in applied work. But, as the articles referenced above demonstrate, the set of indexes *conventionally* reported is much too small to be completely effective.

Item 4 encompasses considerations raised by *EBA*. However, in contrast to the emphasis placed by *EBA* upon sensitivity of point

estimates to a change in the menu of included variables, there is an older tradition of assessing the fragility of models by reference to new data. This is done either through predictive failure, recursive estimation, or interaction with other parts of a model as in simulation analysis. Fragility as an important criterion for model evaluation is therefore not a novel idea. Rather it is the emphasis *EBA* places upon variation in point estimates of a particular coefficient under model respecification which is novel. In fact, an *EBA* would seem to constitute an important part of the evaluative process. It must be a rare instance in which some arbitrariness does not creep into the simplification process, particularly when working with cross-section data. The extreme bounds then provide useful information upon the effects of such arbitrary decisions, at least in respect of the focus coefficient. Such is the way we employ *EBA* in the following case study.

The final category distinguished above challenges a model to encompass or explain alternative models, particularly those originating from past endeavours. Lack of reconciliation between studies is a glaring defect in much current applied research, and this requirement, whether interpreted formally as in Grayham Mizon and Richard (1982), or rather more informally as in James Davidson et al. (1978), must become an essential cornerstone for applied econometric research. Only if it is met can one be truly satisfied that progress has been made in understanding an empirical phenomenon.

In order to contrast the methodology outlined above with the approach of those viewing *EBA* as the cornerstone of econometric work, we will look at the money demand function inquiry presented in Cooley and LeRoy. This paper has been cited approvingly by a number of authors, both for what it said about econometric practice and for its claim about the likely magnitudes of interest elasticities. In doing our comparison we have presumed that the study was meant to be a serious application of the *EBA* methodology, rather than just illustrative. Certainly, there is support for this hypothesis in the stress Cooley and LeRoy laid upon the conclusions drawn from their analysis.

TABLE 2—EXTREME BOUNDS FOR LONG-TERM INTEREST ELASTICITY (*RTB*)

Free Variables	Min	Max
None	-12.14	12.15
<i>RTB</i>	-6.27	2.24
<i>RTB</i> , intercept	-.375	.019

One fact that should by now be apparent from our assignation of *EBA* to a group of methods for model evaluation, is our belief that exclusive attention to the results from it can lead to quite erroneous conclusions about the robustness of parametric inferences. Such tunnel vision tends to distract researchers from the other vital questions needing to be asked. A primary example would be whether the model upon which *EBA* is being practiced is comprehensive enough. Later it is argued that, in Cooley and LeRoy's case, there is ample evidence of it not being so.

For the moment we accept their formulation of the problem, turning instead to one of the items in the list assembled earlier as bedeviling *EBA*; namely, the way in which conclusions on fragility are attendant upon the assumed doubtful/free division. Our Table 2 shows how important such selections were for Cooley and LeRoy's conclusions concerning their second specification (see their Table 2, p. 836).

The extreme bounds shrink dramatically when the intercept is made a free variable. (Note that Cooley and LeRoy, Table 1, p. 835, do not indicate it as doubtful but the bounds of -6.27 and 2.24 from their Table 2 only occur when it is so treated.) With a *t*-statistic of -3.96, such an outcome should not be surprising given our Proposition 1 above. Building a case for the treatment of the intercept as doubtful rather than free would, to our minds, be quite difficult, but the most important lesson from Table 2 is how misleading it is to give the extreme bounds for a single doubtful/free partition of the variables.<sup>5</sup>

<sup>5</sup>In fact, Cooley and LeRoy present a broader range of bounds than those in Table 2, invoking the extra constraint that coefficient estimates must lie in a specified

TABLE 3—ALTERNATIVE ESTIMATES OF THE MONEY DEMAND FUNCTION<sup>a</sup>

	<i>M1</i>	<i>RTB</i>	<i>RSL</i>	<i>INF</i>	<i>GNP</i>	<i>VCC</i>	<i>W</i> <sup>b</sup>
Cooley and LeRoy		.010 (.011)	-.175 (.069)	-.036 (.167)	.372 (.081)	-.009 (.055)	-.107 (.096)
		<i>SEE</i> = .028, <i>D-W</i> = .063					
Simplified Model							
Lag 0		-.003 (.003)	-.111 (.040)	-.156 (.029)	.048 (.051)	-.009 (.018)	.178 (.045)
Lag 1	.866 (.054)	-.005 (.003)	.053 (.043)	-.012 (.026)	.062 (.058)	.007 (.017)	-.178 (.048)
		<i>SEE</i> = .0031, <i>D-W</i> = 1.938, $\rho_1 = .306$ , $\rho_2 = -.219$ , $\rho_3 = .194$					
				(.133)	(.129)	(.131)	
Preferred Model							
	.835 <sup>c</sup> (.047)	-.009 (.002)	-.074 (.021)	-.146 (.026)	.126 (.027)		.178 (.041)
		<i>SEE</i> = .0031, <i>D-W</i> = 2.024, $\rho_1 = .391$ , $\rho_2 = -.301$					
				(.118)	(.112)		

<sup>a</sup>All regressors except the inflation rate are in logs. Constant term is not shown. Standard errors are shown in parentheses, *SEE* = standard deviation of residuals, *D-W* = Durbin-Watson statistic.

<sup>b</sup>For the preferred regression this column is  $\Delta \ln W$ .

<sup>c</sup>Coefficient of lagged real money (in logs).

Table 2 shows that any of the conclusions drawn by Cooley and LeRoy about the magnitude of interest elasticities must be treated with skepticism, even if the output of *EBA* is taken as the dominant source of information on these parameters. The wide bounds relied upon for their critique appear to have been manufactured solely by a particular variable classification. But the inadequacies in their work are even more serious than that. No attention was paid by them at all to the quality of the model used for *EBA*, and it is therefore appropriate that we briefly review it.

In Cooley and LeRoy's specification the demand for real money (*M1*) is held to be a function of two interest rate variables, the savings and loan passbook rate (*RSL*) and the ninety-day Treasury bill rate (*RTB*), real *GNP* (nominal *GNP* divided by the *GNP* deflator, *P*), the current inflation rate (*INF*), the real value of credit card transactions (*VCC*), and real wealth (*W*). They use seasonally adjusted quarterly data for the period 1952:II to 1978:IV, and present (p. 834)

estimates for a loglinear specification. Our estimates of their model are shown in Table 3.<sup>6</sup>

To evaluate Cooley and LeRoy's estimated equation, it is sufficient to note that the most basic index of inadequacy, the Durbin-Watson statistic, is 0.063. This is an example of the situation condemned by C. W. J. Granger and P. Newbold (1974) in which the Durbin-Watson statistic is markedly exceeded by the  $R^2$  and in which arises the danger of the "spurious regression" phenomenon. It is clearly not sensible to investigate fragility with such an inadequate model.

From the above discussion one is entitled to be dubious of the validity of Cooley and LeRoy's claim that the interest elasticity of the demand for money cannot be known with much precision. Nevertheless, it could be correct. Moreover, in view of the prominence of the topic in the literature, and the particular stand taken by Cooley and LeRoy on the issue, it is interesting to see what type

confidence ellipsoid. Those in Table 2 correspond to the 100 percent ellipsoid, and represent wider bounds than most contained in their Table 2.

<sup>6</sup>Cooley kindly made their data available to us. We were able to reproduce their results with the exception that the real wealth elasticity should be  $-0.107$  rather than  $+0.107$ , and the inflation rate should not be in logarithms since negative rates were observed over the sample period.

of model would have eventuated, *given only the data series used by Cooley and LeRoy as input*, if a proper modeling strategy had been followed.<sup>7</sup> That strategy involves the three stages described at the beginning of this section.

#### A. Selection of the General Model

Under the restriction on the universe of available variables, the main direction in which generalization of Cooley and LeRoy's model can take place is in the order of dynamics. Given the wide use of distributed lags in modeling money demand, it seems extraordinary that the authors chose to ignore Christopher Sims' maxim that "a time series regression model arising in econometric research ought in nearly every case to be regarded as a distributed lag model until proven otherwise" (1974, p. 289). The omission of dynamics is even stranger in the light of Cooley and LeRoy's own comments: "Such lagged endogenous variables as the lagged money stock... cannot plausibly be excluded from the demand side either explicitly as observable explanatory variables for the demand for money or implicitly through the time dependence of the error" (p. 840).

Our general model therefore has the same variables as Cooley and LeRoy, but with four lags on all variables (including the dependent). This lag structure seems reasonable considering the data used are quarterly. The period selected for study was, however, shorter than that used by Cooley and LeRoy. John Judd and John Scadding (1982) have recently noted that a large number of studies have experienced difficulty in estimating conventional money demand functions for the post-1973 period. Not only do these models predict poorly, but in a large number of cases such models are dynamically unstable. Various reasons for the poor performances of the models are canvased by Judd and Scadding. Among them, "the most likely

cause of the observed instability in the demand for money after 1973 is innovation in financial arrangements" (p. 1014), which originated from the rapid rise in inflation during the period. In accordance with this view, we restricted ourselves to the subsample 1952:II to 1973:IV, with the first four observations used for constructing up to four lags on all variables.<sup>8</sup>

#### B. Simplification of the General Model

Our first step in simplification of the general model represents an attempt to determine the order of dynamics on each of the variables through a sequence of nested tests. The procedure we use was proposed by Sargan (1980), and has been termed the COMFAC algorithm, due to the fact that it seeks to determine common factors in the distributed lag polynomials associated with each variable. Briefly the logic of the method is as follows.

Suppose the general model had the form

$$(1) \quad y_t = b_1 y_{t-1} + \dots + b_4 y_{t-4} + c_0 x_t + \dots + c_4 x_{t-4} + e_t.$$

With the aid of lag operators, (1) can be rewritten as

$$(2) \quad b(L)y_t = c(L)x_t + e_t,$$

where  $b(L) = 1 - b_1 L - \dots - b_4 L^4$  and  $c(L) = c_0 + c_1 L + \dots + c_4 L^4$  are polynomials in the lag operator  $L$ . If the term  $(1 - \rho_1 L)$  is a common root of both polynomials, (2) can be reexpressed as

$$(3) \quad b^*(L)y_t = c^*(L)x_t + u_t,$$

$$\text{with} \quad b(L) = (1 - \rho_1 L)b^*(L),$$

$$c(L) = (1 - \rho_1 L)c^*(L),$$

$$\text{and} \quad (1 - \rho_1 L)u_t = e_t.$$

<sup>7</sup>The restriction seems necessary to avoid the situation where differences in any conclusions we reach to those of Cooley and LeRoy are simply a consequence of our using information not available to them.

<sup>8</sup>The counterpart to Cooley and LeRoy's model over this shorter period gives parameter estimates  $-2.90$ ,  $-.021$ ,  $-.382$ ,  $-.009$ ,  $-.616$ ,  $-.017$ , and  $-.052$  with standard error of estimate .0083.



TABLE 4—TESTS OF COMMON FACTORS (83 OBSERVATIONS)

Common Factor	Unrestricted Lag Length	Restricted Lag Length	F-Statistic	D.F.	Critical F(0.01)
1,2	4	2	1.755	(12,48)	2.59
3,4	2	0	3.904	(12,60)	2.50
3	2	1	0.426	(6,60)	3.12

An examination of (3) shows that the presence of a common factor has created a new model with maximum lag of three in  $y_t$  and  $x_t$ , and first-order serial correlation ( $AR(1)$ ) in the errors. As there were nine parameters in (2) and only eight in (3), a restriction has been imposed, whose validity may be tested. If the restriction is accepted, the model is capable of being simplified. Moreover, if  $\rho_1$  turns out to be zero, the original model must have had both the orders in  $y_t$  and  $x_t$  overstated.

In our general model there are six regressors apart from the intercept. Hence, in the analogous move from (2) to (3), six restrictions are being imposed in the first attempt at simplification. If the first common factor is accepted, imposing the second leads to a further six restrictions, with the equation error now given as  $AR(2)$ ,  $u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + e_t$ . In this way, each additional common factor restriction leads to six fewer estimated coefficients. Since we have a sequence of nested tests, we set the level of significance of each test at 1 percent so as to have an overall level of significance of approximately 4 percent.<sup>9</sup>

There are two difficulties that can arise in using the  $F$ -test to test the common factors. First, there will generally be multiple minima for the sums of squares (see Sargan, 1980) and, second, there is no guarantee that the common roots in the polynomials attached to the variables are real. In order to guard

against complex roots, we test for two common factors initially, and then test for one common factor only if two are rejected. Thus, in testing for the first two common factors in Table 4, the calculated  $F$ -statistic is 1.7555, the unrestricted (restricted) lag length is 4 (2) and 12 restrictions are being tested (6 associated with each common factor).

Compared with the critical value of  $F(12, 48, .01) = 2.59$ , the calculated statistic is not significant. At this stage, then, the lag length has been reduced from 4 to 2 and the equation error can be expressed as  $AR(2)$ . Testing third and fourth common factors gives a value of 3.904 for the  $F$ -statistic which is significant at 1 percent. Therefore, the third and fourth common factors are rejected. Returning to the second-order lag and testing for the third common factor only gives a value of 0.426 for the  $F$ -statistic. Since the calculated value is significantly less than 3.12, three common factors are accepted. The model can now be expressed as one lag on all variables, with an equation error given as  $AR(3)$ . The resulting model is referred to as the "simplified" one in Table 3.

The following observations are relevant to the simplified dynamic specification given in Table 3. Of the four interest rates, only current  $RSL$  is significant, and, apart from the lagged dependent variable, the only significant lagged variable is real wealth. Moreover, current and lagged wealth have coefficients which add to zero exactly. Neither the current nor lagged real value of credit card transactions exerts a significant effect on real balances. Finally, the third common factor ( $\rho_3$ ) is not significantly different from zero, thereby reducing the implicit lag length of the specification by one.

It is fairly clear that the model is still overparameterized. Accordingly, we imposed a further eight restrictions, namely zero

<sup>9</sup>To test the restrictions, we used the standard  $F$ -test given by  $F = [(\bar{e}'\bar{e} - \hat{e}'\hat{e})/\hat{e}'\hat{e}] \cdot [(T - \hat{k})/r]$ , where  $\bar{e}'\bar{e}$  is the sum of squared residuals from the restricted model,  $\hat{e}'\hat{e}$  is its unrestricted counterpart,  $(T - \hat{k})$  is the degrees of freedom of the unrestricted model, and  $r$  is the number of restrictions to be tested. In this way some allowance is made for the number of parameters estimated in the unrestricted model.

TABLE 5—INDEXES OF ADEQUACY FOR THE PREFERRED MODEL

Statistic Type		Statistic Value		Critical Value	
RESET <sup>a</sup>		3.27		$F(2, 74, .01) = 4.9$	
Diff. Test <sup>b</sup>		2.59		$\chi^2(8, .01) = 20.09$	
Normality Test <sup>c</sup>		1.74		$\chi^2(2, .01) = 9.21$	
Hetero. Test <sup>d</sup>		1.82		$\chi^2(1, .01) = 6.63$	
A.C.F. of squared residuals <sup>e</sup>		1) -.68    3) .51		S.N.D. (.01) = 2.33	
		2) .49    4) .36			
A.C.F. of residuals <sup>f</sup>		1) 1.22    3) 0.73    5) 1.62    7) 0.22			
		2) 1.17    4) 0.53    6) 0.28    8) 1.12			

<sup>a</sup>The  $F$ -test that the coefficients of the predictions squared and cubed in the regression of the residuals against these and the derivatives are zero. Computation was done via partitioned inversion to avoid serious numerical inaccuracy.

<sup>b</sup>The differencing test of Charles Plosser et al. (1982). One iteration of Sargan's (1959)  $AIV$  estimator upon the differenced model was performed from the estimates in Table 3. Instruments for the derivatives with respect to the coefficients of  $M_{t-1}$  and  $u_{t-1}$  were constructed as in Plosser et al. (fn. 7).

<sup>c</sup>The joint normality test of K. O. Bowman and L. R. Shenton (1975), or Anil Bera and Carlos Jarque (1981).

<sup>d</sup>The  $LM$  test that  $\gamma = 0$  in  $\sigma^2 = \sigma^2(E(y_t))^\gamma$  where  $y_t$  is the dependent variable of a regression. Pagan et al. (1981) derive this  $LM$  test but it was proposed originally as a test for heteroscedasticity by F. J. Anscombe (1961).

<sup>e</sup>The  $t$ -statistics were formed by regressing the squared residuals against their lagged values. This approach was used by Granger and Allan Andersen (1978) for the detection of nonlinear models but can also be used to check for Engle's (1982a)  $ARCH$  effects or as a general specification error test.

<sup>f</sup>Writing the model as a nonlinear regression  $y = f(X; \theta) + \varepsilon$ , the  $t$ -statistics that the coefficient of the lagged residuals  $\hat{\varepsilon}_{t-j}$  are zero in the regressions of  $\hat{\varepsilon}_t$  against  $\hat{\varepsilon}_{t-j}$  and  $\partial f_t / \partial \theta$  for  $j = 1, \dots, 8$ .

coefficients for  $VCC$  and the lagged values of  $RTB$ ,  $RSL$ ,  $INF$ ,  $GNP$ , and  $VCC$ , a zero sum for the wealth coefficients, and a zero value for the third common factor. The calculated  $F$ -statistic of 1.001 is significantly less than  $F(8, 66, 0.01) = 2.8$ , leading to acceptance of the restrictions. Our preferred model is therefore the last one listed in Table 3.

### C. Quality Control: Is the Model a Lemon?

How does the estimated model in Table 3 stand up to the five criteria for quality control listed at the beginning of this section? With the exception of the term  $\Delta \ln W_t$ , it constitutes a very traditional specification of money demand. The presence of the change in, rather than the level of, wealth is however consistent with theoretical considerations. If transactions requirements are held constant, that is,  $GNP_t$  is fixed, the fact that money ( $M1$ ) is an asset dominated for portfolio purposes by interest-bearing deposits of near equal liquidity suggests that the long-term

wealth effect should be zero. In the short run though, it has been frequently noted that changes in wealth are initially held as demand deposits before reallocation, and the combination of  $\Delta \ln W_t$  and the lagged dependent variable describes such a process, the implied lag distribution being .178, -.029, -.025, etc. Perhaps the only difficulty with such an interpretation is that the portfolio reallocation process is not faster.

Table 5 investigates whether there are any obvious "inferential monsters lurking beyond the horizon," by augmenting the moments of the preferred model with a number of variables designed to capture inadequacy. No striking deficiencies are in evidence. A number of other experiments were conducted to determine whether it was possible to reject the chosen model by the addition of particular variables. These included a number of lags in real  $GDP$ ,  $RSL$ , etc., time trend, seasonal dummies, and estimation with up to seventh-order serial correlation pattern. None of these augmentations was found to contribute anything of significance. A final point

worth mentioning is that *t*-statistics, made robust to heteroscedasticity as suggested by Halbert White (1980), were about 10 percent higher than those in Table 3. The only exception to this rule—that for inflation—was only slightly smaller.

A check on parameter constancy is available by examining the size of prediction errors made when an equation is estimated over a particular sample and then used to forecast out of sample.<sup>10</sup> In this vein, the preferred model was estimated to 1970:IV and one-step prediction errors were generated from 1971:I to 1973:IV by augmenting the preferred equation with the Type *B* constructed variables in Pagan and Desmond Nicholls (1984). The "*F*-test" that the coefficients on the twelve constructed variables were jointly zero was 1.58, well below the critical *F*(12, 62, .01) value of 2.49. Although an examination of the individual errors does reveal one large error, namely that for 1972:I, where the *t*-value was 2.58, the prediction errors for 1971–73 were much the same as the sample errors, with an average absolute value of .4 percent.<sup>11</sup>

#### D. Are Interest Elasticities of Money Demand Zero?

As our model was of satisfactory quality to 1973:IV, it is reasonable to utilize it to shed light on the question of whether data is uninformative about interest elasticities, as alleged by Cooley and LeRoy.<sup>12</sup> Conditional

upon the structure of the final model being valid, we can say that all variables in the estimated relationship (including *both* interest rates) are highly significant, and to adopt their hypothesis of a zero interest rate effect as an acceptable interpretation of the data would be totally inappropriate. To be sure, this final specification was arrived at after a decision in which an arbitrary group of variables was dropped because of insignificance. To assure readers that the well-defined interest elasticities found in our preferred model were not dependent upon this action, and to illustrate what we believe is the place of *EBA*, we computed the extreme bounds for the two long-run interest elasticities. This was done by making the coefficients on either *RSL* or *RTB* the focus, treating all excluded variables as doubtful, and using the estimates of parameters on  $\ln M_{t-1}$ ,  $\ln RTB_t$  (or  $\ln RSL_t$ ),  $\ln RTB_{t-1}$  (or  $\ln RSL_{t-1}$ ) associated with the bounds to obtain long-run responses. To be consistent with Cooley and LeRoy, we concentrate upon the long-run elasticities as they summed lagged coefficients when dynamics were admitted. These bounds were extremely narrow, being  $-.053$  to  $-.068$  (*RTB*) and  $-.400$  to  $-.441$  (*RSL*), indicating that the effect of interest rates upon money demand was not sensitive to our decision to exclude certain variables.

#### V. Conclusions

That applied econometrics is not currently in the most robust of health is hard to deny, and it would be difficult to find as entertaining or as perceptive an analysis of its ills as that found in Leamer's various articles. What concerns us is that the prescriptions made in those articles are inappropriate, in part because of faulty diagnosis. Extreme bounds analysis (*EBA*) is most emphatically *not* the medicine to cure an ailing patient.

Section I argued that extreme bounds are generated by the imposition of highly arbitrary

<sup>10</sup>A more detailed analysis is available in our earlier working paper.

<sup>11</sup>Although our model gave satisfactory performance up to 1973, just like automobiles, age finally caught up with it, and after that date its predictive performance declined dramatically. For the twelve quarters after 1973:IV, the *F*-test that prediction errors were zero was 5.69, with only the errors for 1974 not being significantly different from zero individually. The absolute error was 1.7 percent over this three-year period. Thus Stephen Goldfeld's (1976) puzzle of the "missing money" is certainly not resolved by working with Cooley and LeRoy's data alone.

<sup>12</sup>Encompassing tests were also advocated to assess model quality. These are not really possible here given the restriction placed upon the data set, although it is clear that our model dominates those which exclude either *RSL* or *RTB*, the inflation rate, wealth or

explicit dynamics. In Hendry and Richard's terminology, our model strongly variance-encompasses Cooley and LeRoy's as is apparent from the standard errors of estimate in our fn. 8 and Table 3.

trary, and generally unknown, restrictions between the parameters of a model. Exactly why such bounds should be of interest therefore becomes something of a mystery. Furthermore, as shown in Sections II and III, the methodology is flawed on other grounds. *EBA* demands a general, adequate model from which the bounds may be derived, and a consensus over which variables are critical to a relationship. These are highly questionable conventions and we demonstrated, both theoretically and empirically, that deviations from them almost completely negate the utility of *EBA*.

After largely rejecting *EBA*, Section IV of the paper moved on to our own diagnosis and prescription. Both are founded on the belief that many of the difficulties applied econometrics currently faces originate in the very poor attempts currently made to accurately describe the process whereby a model was selected, and to ascertain its adequacy. Acceptance of this proposition leads to the necessity for the establishment and promulgation of standards with which to conduct applied research. Many other disciplines have faced and taken steps to solve this problem, and movement in this direction is long overdue in econometrics. With these considerations in mind we proposed a three-stage approach to modeling, involving the selection and subsequent simplification of a general model and a rigorous evaluation of any preferred model. Under the latter heading, five ways of performing such an evaluation were distinguished. It may not be too fanciful to think of such criteria as a "checklist" to be applied when reviewing or performing applied work. Only if a model passes most items on the list should it be seriously considered as augmenting our knowledge.

Having set up some yardsticks with which to evaluate models, Section IV applied them to the money demand example in Cooley and LeRoy. Their specification was found to fail even the simplest of these criteria, making any conclusions drawn from it highly suspect. In sharp contrast to this failure, the application of a modeling strategy beginning with a general model and progressively constraining the parameter space led to a representation which passed all items of the

checklist. This example highlighted the benefit of a systematic approach to modeling and model evaluation.

In closing, a confession. We are only too aware that what has been described are the necessary rather than sufficient conditions for taking the con out of econometrics. As any users of corporate accounts will be aware, there are many ways around standards. But that is not to deny their value. It serves only to highlight the need.

## REFERENCES

- Anscombe, F. J., "Examination of Residuals," *Proceedings, Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1961, 4, 1-36.
- Bera, Anil K. and Jarque, Carlos M., "An Efficient Large Sample Test for Normality of Observations and Regression Residuals," *Working Papers in Economics and Econometrics*, No. 040, Australian National University, 1981.
- Bowman, K. O. and Shenton, L. R., "Omnibus Contours for Departures from Normality Based on  $\sqrt{b_1}$  and  $b_2$ ," *Biometrika*, No. 2, 1975, 62, 243-50.
- Christ, Carl F., "Econometrics in Economics: Some Achievements and Challenges," *Australian Economic Papers*, December 1967, 6, 155-70.
- Cooley, Thomas F., "Specification Analysis with Discriminating Priors: An Application to the Concentration Profits Debate," *Econometric Reviews*, No. 1, 1982, 1, 97-128.
- and LeRoy, Stephen F., "Identification and Estimation of Money Demand," *American Economic Review*, December 1981, 71, 825-44.
- Davidson et al., James E. H., "Econometric Modelling of the Aggregate Time-Series Relationship Between Consumers' Expenditure and Income in the United Kingdom," *Economic Journal*, December 1978, 88, 661-92.
- Dhrymes, Phoebus J., "Comment," *Econometric Reviews*, No. 1, 1982, 1, 129-32.
- Dicks-Mireaux, Louis and King, Mervyn, "Pen-sion Wealth and Household Savings: Tests

- of Robustness," *Journal of Public Economics*, February/March 1984, 23, 115-39.
- Engle, Robert F., (1982a) "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, July 1982, 50, 987-1007.
- \_\_\_\_\_, (1982b) "A General Approach to Lagrange Multiplier Model Diagnostics," *Journal of Econometrics*, October 1982, 20, 83-104.
- Fiebig, D. G., "A Bayesian Analysis of Inventory Investment," *Empirical Economics*, 1981, 6, 229-37.
- Goldfeld, Stephen M., "The Case of the Missing Money," *Brookings Papers on Economic Activity*, 3: 1976, 683-730.
- Granger, C. W. J. and Andersen, A., *An Introduction to Bilinear Time Series Models*, Gottingen: Vandenhoeck and Ruprecht, 1978.
- \_\_\_\_\_, and Newbold, P., "Spurious Regressions in Econometrics," *Journal of Econometrics*, July 1974, 2, 111-20.
- Hendry, David F., "Econometrics: Alchemy or Science?," *Economica*, November 1980, 47, 387-406.
- \_\_\_\_\_, "Econometric Modelling: The Consumption Function in Retrospect," *Scottish Journal of Political Economy*, 1983, 30, 193-220.
- \_\_\_\_\_, and Richard, Jean-François, "On the Formulation of Empirical Models in Dynamic Econometrics," *Journal of Econometrics*, October 1982, 20, 3-33.
- Judd, John P. and Scadding, John L., "The Search for a Stable Money Demand Function: A Survey of the Post-1973 Literature," *Journal of Economic Literature*, September 1982, 20, 993-1023.
- Leamer, Edward E., *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, New York: Wiley & Sons, 1978.
- \_\_\_\_\_, "SEARCH, A Linear Regression Computer Package," mimeo., University of California-Los Angeles, 1981.
- \_\_\_\_\_, "Sets of Posterior Means with Bounded Variance Priors," *Econometrica*, May 1982, 50, 725-36.
- \_\_\_\_\_, "Let's Take the Con Out of Econometrics," *American Economic Review*, March 1983, 73, 31-43.
- \_\_\_\_\_, and Leonard, Herman, "Reporting the Fragility of Regression Estimates," *Review of Economics and Statistics*, May 1983, 65, 306-17.
- Mizon, Grayham E. and Richard, Jean-François, "The Encompassing Principle and Its Application to Non-nested Hypotheses," paper presented to the European meeting of the Econometric Society, Dublin 1982.
- Pagan, A. R. and Hall, A. D., "Diagnostic Tests as Residual Analysis," *Econometric Reviews*, No. 2, 1983, 2, 159-218.
- \_\_\_\_\_, \_\_\_\_\_, and Trivedi, P. K., "Assessing the Variability of Inflation," Working Papers in Economics and Econometrics, No. 049, Australian National University, 1981.
- \_\_\_\_\_, and Nicholls, D. F., "Estimating Predictions, Prediction Errors and their Standard Deviations Using Constructed Variables," *Journal of Econometrics*, March 1984, 24, 293-310.
- Plosser, Charles I., Schwert, G. William and White, Halbert, "Differencing as a Test of Specification," *International Economic Review*, October 1982, 23, 535-52.
- Ramsey, James B., "Perspective and Comment," *Econometric Reviews*, No. 2, 1983, 2, 241-48.
- Sargan, J. D., "The Estimation of Relationships with Autocorrelated Residuals by the Use of Instrumental Variables," *Journal of the Royal Statistical Society, Series B*, No. 1, 1959, 21, 91-105.
- \_\_\_\_\_, "Wages and Prices in the United Kingdom: A Study in Econometric Methodology," in P. E. Hart et al., eds., *Econometric Analysis for National Economic Planning*, London: Butterworths, 1964, 25-63.
- \_\_\_\_\_, "Some Tests of Dynamic Specification for a Single Equation," *Econometrica*, May 1980, 48, 879-97.
- Sims, Christopher A., "Distributed Lags," in M. D. Intriligator and D. A. Kendrick, eds., *Frontiers of Quantitative Economics*, Vol. II, Amsterdam: North-Holland, 1974, 289-338.
- White, Halbert, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, May 1980, 48, 817-38.

# Sensitivity Analyses Would Help

By EDWARD E. LEAMER\*

A fragile inference is not worth taking seriously.

All scientific disciplines routinely subject their inferences to studies of fragility. Why should economics be different? It hasn't been different up to now. Nor do I think it ever will be, notwithstanding the comments of Michael McAleer, Adrian Pagan, and Paul Volker (1985).

Decentralized studies of fragility are common whenever an inference matters enough to attract careful scrutiny. When Isaac Erlich (1975) claims to have demonstrated that capital punishment deters murders, he elicits a great outpouring of papers that show how the result depends on which variables are included (B. Forst 1977), which observations are included (A. Blumstein et al., 1978), how simultaneity problems are dealt with (P. Passell, 1975), etcetera, etcetera. These disorganized studies of fragility are inefficient, haphazard, and confusing.

What we need instead are organized sensitivity analyses. We must insist that all empirical studies offer convincing evidence of inferential sturdiness. We need to be shown that minor changes in the list of variables do not alter fundamentally the conclusions, nor does a slight reweighting of observations, nor correction for dependence among observations, etcetera, etcetera.

I have proposed a form of organized sensitivity analysis that I call "global sensitivity analysis" in which a neighborhood of alternative assumptions is selected and the corresponding interval of inferences is identified. Conclusions are judged to be sturdy only if the neighborhood of assumptions is wide enough to be credible and the corresponding interval of inferences is narrow

enough to be useful. But when an incredibly narrow set of assumptions is required to produce a usefully narrow set of conclusions, inferences from the given data set are reported to be too fragile to be believed.

In dramatic conflict with real data analyses, theoretical econometricians behave as if a given data set admitted a unique inference. This priesthood takes as their self-appointed task the uncovering of the elaborate method by which the unique inference can be squeezed from a data set. Indeed, this is the reaction of McAleer et al., who offer a method of squeezing Thomas Cooley and Stephen Leroy's (1981) data set. They propose to deal with specification ambiguity by charting one *ad hoc* route through the thicket of possible models. Complicated *ad hoc* searches like the one they suggest have no support in statistical decision theory, and virtually none in classical sampling theory. What is to be made of a procedure that sets scores of parameters to zero if they are not "statistically significant" at arbitrarily chosen levels of significance? And what inferences are allowable after a model passes a battery of "specification error" tests that are sometimes more numerous than even the set of observations? This recommendation of McAleer et al. merits the retort: "There are two things you are better off not seeing in the making: sausages and econometric estimates," to which they might reply: "It must be right, I've been doing it since my youth."

## I. Extreme Bounds Analysis

McAleer et al. direct their criticisms at what they call "extreme bounds analysis."<sup>1</sup>

<sup>1</sup>I prefer to use the words "extreme bounds" to refer to the largest possible set of inferences that a given data set will admit. This largest possible set of inferences depends on the largest possible family of models cum priors that the researcher is willing to entertain. The family of models that give rise to what McAleer et al.

\*Department of Economics, University of California, Los Angeles, CA 90024. Helpful comments from Sebastian Edwards, John Riley, and especially Harold Demsetz are gratefully acknowledged.

Since most of their concerns stem from misunderstandings about the setting in which this applies, I will begin my reply with a careful description of the specification ambiguity that properly gives rise to this form of sensitivity analysis. This setting is most accurately described in Bayesian terms, but to communicate with non-Bayesian readers I often adopt the more familiar language of the specification searcher. I fear it is the translation that causes the problems.

In Bayesian terms, the "extreme bounds" are applicable when the prior distribution for a subset of coefficients is located at the origin but is otherwise unspecified, and the prior distribution for the other coefficients is "diffuse." A sensitivity analysis is then performed to determine if features of the posterior distribution depend importantly on the way this partially defined prior distribution is fully specified. It is particularly easy to search over the set of alternative posterior distributions to find the extreme posterior modes of linear combinations of coefficients, ergo "extreme bounds."

It has been my experience that this Bayesian description of the "extreme bounds analysis" meets with equally small amounts of approval and understanding. To combat this ignorance and suspicion, I have tried to market the idea in the following disguise. Imagine the estimation of a regression in a setting in which there are a few variables that are always left in the equation (the free variables), and some others that the researcher feels comfortable experimenting with (the doubtful variables).<sup>2</sup> Normally, this experi-

mentation is limited to a small subset of the possible models that could have been estimated. Suppose instead that we consider the whole continuum of models in which the free variables and any one linear combination of the doubtful variables are included in the model. If it turns out that inferences about issues of interest are essentially the same for all choices of the linear combination of doubtful variables, then there need be no debate about which doubtful variables ought to be deleted. If, as is often the case, this bound turns out to be uselessly wide, then either the inferences are reported to be too fragile to be useful, or the bounds are narrowed in one way or another.

This disguised treatment of the sensitivity issues is quite alright up to this point, but an explicitly Bayesian framework is required to discuss sensibly any attempt to narrow the bounds. Unfortunately, the transition to the proper Bayesian foundation of the analysis is not easy for many people. For example, McAleer et al., and others as well, remark that the extreme estimates come from a model that includes a very strange linear combination of the doubtful variables. To cure this "problem," it is often proposed to restrict the bounds to the set of regressions that exclude subsets of the doubtful variables, a procedure that can be called "all subsets regression." But this advice depends on the parameterization of the model, which is usually chosen by whim rather than design. In a distributed lag model, the "all subsets regressions" would be different if  $x(t)$  and  $x(t-1)$  are doubtful variables than if  $x(t)$  and  $x(t) - x(t-1)$  are doubtful variables, since in the former case the omission of  $x(t)$  leaves  $x(t-1)$  as the included variable, whereas in the latter case  $x(t) - x(t-1)$  is retained.<sup>3</sup>

How should the parameters be defined? The subject of sensitivity analysis provides an answer. Regression coefficients should be defined about which prior opinions are inde-

call "extreme bounds analysis" may or may not be the widest set under consideration. In order to bring attention to this difference in language, I will put quotation marks around the words when used in the sense of McAleer et al.

<sup>2</sup>A terminological error that was first made in my earlier study (1978, p. 194) and adopted vigorously by myself and H. B. Leonard (1983) is the use of "focus" rather than "free" to describe the variables that are always in the equation. It is often but not always the case that the free variables are also the focus variables. An earlier version of the paper by McAleer et al. had comments that rested on this terminological error, and at my suggestion they have adopted the word free.

<sup>3</sup>As a matter of fact, unless you are prepared to commit to a particular coordinate system, the restriction to all subsets regression is vacuous since all subsets in all coordinate systems will reproduce exactly the extreme bounds.

pendent. If, before the data are observed, opinions about  $b(1)$  do not depend on information about  $b(2)$ , then these are suitable choices for the pair of coefficients but  $b(1)$  and  $b(1) + b(2)$  are not. The reason for defining the model with a priori independent coefficients is that, if the prior distribution cannot be more fully specified, the "all subsets" bound applies (see my article with Gary Chamberlain, 1976) and is generally narrower than the "extreme bounds." For the distributed lag example, the model should be written as  $y = b(1) \times (x(t) + x(t-1)) + b(2) \times (x(t) - x(t-1))$  if the best guess of  $b(1)$ , the steady-state responsiveness of the dependent variable, is unaffected by information about  $b(2)$ , the responsiveness to changes in the level of the stimulus. In that event, if both variables are regarded to be doubtful, it makes sense to see what happens to the estimate of  $b(1)$  when  $b(2)$  is set to zero, and it makes sense to see what happens to the estimate of  $b(2)$  when  $b(1)$  is set to zero.

Classical inference offers no advice on how to choose a parameterization. It is altogether immaterial whether  $x(t)$  and  $x(t-1)$  are the explanatory variables, or  $x(t)$  and  $x(t) - x(t-1)$ , or  $(x(t) + x(t-1))$  and  $(x(t) - x(t-1))$ . As a result, there is total confusion in the literature about anything that depends on the coordinate system, the multicollinearity problem being the prime example.<sup>4</sup>

"All subsets" regression replaces the extreme bounds if the prior distribution is suitably restricted. There are many other bounds that could be computed depending on how fully the prior distribution is specified. The restricted family of prior distributions that I often find appealing has a fixed prior mean

and an "interval" of prior covariance matrices (see my 1982 article). The "extreme bounds analysis" is one special case with an unbounded interval of covariance matrices for a subset of coefficients and a sharply defined covariance matrix for the others. This somewhat unusual interval of prior distributions is often and properly criticized for being too wide on one subset and too narrow on the other. Indeed that is the criticism of McAleer et al., though of course they don't use this language to express their concerns.

McAleer et al. offer some useful caveats about "extreme bounds analysis," pointing out the important point that the bounds depend on the family of prior distributions. I am confident that on reflection the properties that seem to bother them will be judged to be altogether desirable, and I welcome the glare of publicity that they offer. These properties are:

1) The bounds depend on the choice of variables that are treated as doubtful.

2) If coefficients are set to values other than zero, different bounds will result. (This amounts to saying that the bounds depend on the location of the prior distributions.)

3) The bounds for the free coefficients will be wider the more statistically significant are the doubtful variables.

4) If a variable is treated as doubtful, a zero estimate for its coefficient is necessarily obtainable.

5) If there are two or more doubtful variables, a coefficient on a doubtful variable may be either positive or negative, regardless of the degree of correlation between the variables.

If you do find these properties undesirable, it must be that you are implicitly rejecting the family of prior distributions on which they are based. It is no surprise that "extreme bounds analysis" does not apply in all cases. As a matter of fact, it rarely applies. I use it primarily as a warm-up device for introducing the kind of bounds that I think are truly applicable. I welcome this kind of criticism of the "extreme bounds analysis" since it affords me the opportunity to enrich the vocabulary of the conversation by demonstrating the proper Bayesian foundation of the analysis. With this enriched vocabulary it

<sup>4</sup>After all, there is always a coordinate system in which the "variables" are orthogonal and in which by traditional standards there is no multicollinearity problem. I have repeatedly but unsuccessfully tried to explain this simple point. Expressed most dramatically, since orthogonality is a happenstance of the coordinate system, there may still be a multicollinearity problem if the explanatory variables are orthogonal, though it is more accurate to say that there is a problem of dimensionality. For more, see my earlier articles (1973 or 1978, or 1983).



is possible to demonstrate how other, more relevant bounds may be computed.

## II. Global Sensitivity Analysis

The real point of disagreement between myself and other econometricians is that I believe the only "model selection" game in town ought to be the global sensitivity game. Except for issues associated with data-instigated models and simplification problems, both of which are dealt with in my earlier study (1978), a sensible and general characterization of the problem of inference begins with a broad family of alternative models and a representative, but hypothetical, prior distribution over that family. Because no prior distribution can be taken to be an exact representation of opinion, a global sensitivity analysis is carried out to determine which inferences are fragile and which are sturdy. A neighborhood of prior distributions around the representative distribution is selected and inferences that depend in a significant way on the choice of prior from this neighborhood are judged to be fragile. Ideally, the neighborhood of distributions is credibly wide, and the corresponding interval of inferences is usefully narrow. But if it is discovered that an incredibly narrow neighborhood of prior distributions is required to produce a usefully narrow set of inferences, then inferences from the given data set are suspended, and pronounced too fragile to serve as a basis for action.

I don't pretend that this research strategy is easy. It certainly is not a comfortable one for those trained in and wedded to classical inference. It isn't something that can be done by a computer without the aid of a human. I recognize that it will take some considerable experience with these methods until we can decide what constitutes a "credibly" wide set of assumptions. One thing that is clear is that the dimension of the parameter space should be very large by traditional standards. Large numbers of variables should be included, as should different functional forms, different distributions, different serial correlation assumptions, different measurement error processes, etcetera, etcetera.

In principle, a global sensitivity study should be carried out with respect to all dimensions of the model in one grand exercise. Alas, the mathematical/computational problems in dealing with the list of etceteras are very severe. But since the longest journey begins with a single step, a piecemeal approach is proposed in which the sensitivity analysis is carried out with respect to a limited number of dimensions of the model. The parameters that lend themselves to the most congenial analysis are regression coefficients, but I can give you some interesting and useful results on errors in variables (Steven Klepper and myself, 1984), and on the distribution of residuals (my 1981, 1983a, 1984 articles, and C. Z. Gilstein and myself, 1983a; b).

When you review a global sensitivity analysis, you need ask yourself two important questions: is the dimension of the model space adequate? Is the neighborhood of prior distributions the right width—wide enough to include all sensible distributions but not so wide that it includes nonsensical ones? I interpret the comments on Cooley and Leroy by McAleer et al. to amount to a healthy discussion of exactly these two issues. This I believe is precisely the form that debates about empirical results ought to take. McAleer et al. find Cooley and Leroy's space of models inadequate because it includes no parameters for the dynamics, and they find the neighborhood of priors to be both too wide (allowing funny linear restrictions) and too narrow (some of the free variables might be doubtful). The very fact that they are able to make these comments reveals the value of the global sensitivity framework for focusing the issues.

McAleer et al.'s "most serious" charge is:

Unless extreme bounds are presented for *all* possible classifications of variables as doubtful and free, an observer cannot be certain that the selection does not constitute a "con job." *Selectivity in regression reporting therefore has as an exact analog in EBA the different classifications of variables as doubtful and free.* [p. 298]

This charge seems to be based on the belief that the distinction between free and doubtful variables is altogether arbitrary. Actually the split should be selected to represent as accurately as possible the other relevant information that is required to draw sensible inferences from the given data set. When readers differ concerning their willingness to believe and/or to use other relevant information, then a menu of inferences should be presented, and as clear as possible a statement should be made about the assumptions that are necessary to make one inference or another. When the menu isn't broad enough to suit your tastes, there is no reason to believe the inferences claimed by the author. For example, McAleer et al. find the menu of inferences offered by Cooley and Leroy to be uninteresting because the models include no parameters for dynamics. McAleer et al. seem not to have been "conned." Why are they worried about the intelligence of the rest of us? Thus I think a global sensitivity analysis is neutral at worst with respect to dishonesty. After all, the finding that a data set does not admit a sturdy inference is news worthy of publication. On the other hand, current institutions clearly encourage, and have produced, either delusion or deceit.

### III. A Sugar Pill for a Nearly Terminal Patient

An epidemic of overparameterization debilitates our data analyses. We need strong medicine to combat this disease. I know a global sensitivity analysis is a bitter pill to swallow. But try it, I think it's going to make us all feel much better. Maybe not entirely well, but better anyway. The sugar pill that McAleer et al. offer has the pleasant taste of the familiar, but past experience suggests that it won't do any good.

Global sensitivity analysis can deal in principle with all varieties of parameters. "Extreme bounds analysis" is one example of a global sensitivity analysis. In order to make clear how the proposal of McAleer et al. compares with "extreme bounds analysis," we need to focus on the problem for which the "extreme bounds analysis" is intended, namely the choice of variables. The comments that follow apply equally well to

other parameter spaces, since the competing methodological approaches are the same regardless of the nature of the statistical assumptions selected by different parameters. In particular, the complaint that "extreme bounds analysis" doesn't deal with serial correlation, nonnormality, etcetera, is quite irrelevant. After all, it is hardly reasonable to complain that brain surgery can't cure a hangnail.

What McAleer et al. propose for the problem of choice of variables is a combination of backward and forward step-wise (better known as unwise) regression. The variables are divided into three groups, say  $x(1)$ ,  $x(2)$ , and  $x(3)$ . The first equation that is estimated includes all the variables from the sets  $x(1)$  and  $x(2)$  and excludes all the variables from the set  $x(3)$ . This "general model" is subjected to a sequence of tests to determine if subsets of the variables comprising  $x(2)$  can be omitted. The order for imposing the restrictions and the choice of significance level are arbitrary, though in some cases set by convention. Thus results a "preferred model." Then this preferred model is subjected to a battery of "specification error" tests to determine if variables in the subset  $x(3)$  should be included. Again the order and the levels of significance are arbitrary. So is the split into the subsets  $x(1)$ ,  $x(2)$ , and  $x(3)$ . What meaning should be attached to all of this?

### REFERENCES

- Blumstein, A., Cohen, J. and Nagin, D., *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates*, Washington: National Academy of Sciences, 1978.
- Chamberlain, Gary and Leamer, Edward E., "Matrix Weighted Averages and Posterior Bounds," *Journal of the Royal Statistical Society*, No. 1, 1976, 38, 73-84.
- Cooley, Thomas F. and Leroy, Stephen F., "Identification and Estimation of Money Demand," *American Economic Review*, December 1981, 71, 825-44.
- Erlich, Isaac, "The Deterrent Effect of Capital Punishment: A Question of Life and Death," *American Economic Review*, June

- 1975, 65, 397-417.
- Forst, B., "The Deterrent Effect of Capital Punishment: A Cross State Analysis of the 1960's," *Minnesota Law Review*, May 1977, 61, 743-67.
- Gilstein, C. Z. and Leamer, E. E., (1983a) "Robust Sets of Regression Estimates," *Econometrica*, March 1983, 51, 321-33.
- \_\_\_\_\_, and \_\_\_\_\_, (1983b) "The Set of Weighted Regression Estimates," *Journal of the American Statistical Association*, December 1983, 78, 942-48.
- Klepper, Steven and Leamer, Edward E., "Consistent Sets of Estimates for Regressions with Errors in All Variables," *Econometrica*, January 1984, 52, 163-83.
- Leamer, Edward, E., "Multicollinearity: A Bayesian Interpretation," *Review of Economics and Statistics*, August 1973, 55, 371-80.
- \_\_\_\_\_, *Specification Searches*, New York: Wiley & Sons, 1978.
- \_\_\_\_\_, "Sets of Estimates of Location," *Econometrica*, January 1981, 49, 193-204.
- \_\_\_\_\_, "Sets of Posterior Means with Bounded Variance Priors," *Econometrica*, May 1982, 50, 726-36.
- \_\_\_\_\_, (1983a) "Let's Take the Con Out of Econometrics," *American Economic Review*, March 1983, 73, 31-43.
- \_\_\_\_\_, (1983b) "Model Choice and Specification Analysis," in Zvi Griliches and Michael D. Intriligator, eds., *Handbook of Econometrics*, Vol. I, Amsterdam: North-Holland, 1983, 285-330.
- \_\_\_\_\_, "Global Sensitivity Results for Generalized Least Squares Estimates," *Journal of the American Statistical Association*, December 1984, 79, 867-70.
- \_\_\_\_\_, and Chamberlain, Gary, "A Bayesian Interpretation of Pretesting," *Journal of the Royal Statistical Society*, No. 1, 1976, 38, 85-94.
- \_\_\_\_\_, and Leonard, H. B., "Reporting the Fragility of Regression Estimates," *Review of Economics and Statistics*, May 1983, 65, 306-17.
- McAleer, Michael, Pagan, Adrian R. and Volker, Paul A., "What Will Take the Con Out of Econometrics?," *American Economic Review*, June 1985, 75, 293-307.
- Passell, P., "The Deterrent Effect of the Death Penalty: A Statistical Test," *Stanford Law Review*, November 1975, 28, 61-80.

# An Experimental Test of the Consistent-Conjectures Hypothesis

By CHARLES A. HOLT\*

A common way of analyzing multiperiod oligopoly models without dynamic interactions in the payoff structure is to compute a Nash equilibrium for each period taken separately. Many economists believe that behavior in a repeated market game cannot be predicted accurately with a period-by-period sequence of such "static" Nash equilibria, but an explicitly dynamic analysis can be extremely difficult unless the class of feasible dynamic strategies is restricted.<sup>1</sup>

There is an embarrassing multiplicity of alternative oligopoly "solutions" that are computationally less complex than game-theoretic approaches to multiperiod games. Many of these alternative solutions can be classified as conjectural variations models in which firms are assumed to conjecture that changes in their own decisions will induce reactions by other firms. These reactions are typically assumed to be characterized by

functions that are locally linear. Almost any configuration of decisions can be an equilibrium for some conjectured reaction functions, so these models have little empirical content unless the reaction functions themselves are determined endogenously.

Timothy Bresnahan (1981) has proposed a consistency condition that can often be used to determine specific conjectured reactions. Martin Perry provides a clear explanation of this consistency condition in the context of a duopoly in which firms' decisions are output quantities:

Each firm's first-order condition defines its profit-maximizing output as a reaction function on (1) the output of the other firm and (2) the conjectural variation about the other firm's response. Thus a conjectural variation by one firm about the other firm's response is consistent if it is equivalent to the derivative of the other firm's reaction function with respect to the first firm's output at equilibrium.

[1982, p. 197]

\*Department of Economics, University of Virginia, Charlottesville, VA 22901. This research was funded by the National Science Foundation. Laura Cohen, Brad Hauck, and Anne Villamil assisted in administering the experiments. Peggy Claytor assisted in the preparation of this manuscript. I am grateful to Dan Alger, Alfonso Novales, Robert Porter, Roger Sherman, and Joel Slemrod for comments on an earlier draft.

<sup>1</sup>James Friedman (1977) discusses the existence of Nash equilibria in a general class of reaction function strategies, but one cannot actually compute nondegenerate equilibrium reaction functions for even the simplest quadratic payoff structures. More severe restrictions on the strategy spaces can produce results. For example, Richard Cyert and Morris DeGroot (1970) use backward induction to compute Nash equilibrium sequences of outputs for a finite horizon duopoly model in which firms make output decisions in alternate periods. Friedman's "balanced temptation equilibrium" is a Nash equilibrium for a supergame in which firms choose contingent strategies that specify an equilibrium output level and a commitment to a permanent switch to the firm's static Cournot output if another firm increases its output above its equilibrium level. Edward Green and Robert Porter (1984) have analyzed a stochastic generalization of this balanced temptation equilibrium.

Many economists have found this notion of consistency to be appealing; Perry cites a large number of recent working papers on the theoretical properties of consistent-conjectures equilibria.

Although not explicitly dynamic, the consistent-conjectures equilibrium (CCE) approach initially seemed plausible to me because it predicts deviations from a static Nash equilibrium that are qualitatively consistent with the data reported in several published laboratory experiments with student subjects. These experiments, however, were not designed to provide a clear distinction between the CCE and other equilibrium concepts. This paper reports the results of an experiment designed specifically to test the consistent-conjectures hypothesis.

In Section I, the computation of a consistent-conjectures equilibrium is explained in the parametric context that is used to construct the experiment. Section II contains a discussion of how the payoff structures used in the previous laboratory experiments must be modified to permit a good test of the consistent-conjectures hypothesis. In Section III, I report the results of an experiment in which the theoretical predictions of the static Nash and consistent-conjectures equilibria are quite different. The data are clearly inconsistent with the CCE hypothesis. A related experiment is discussed in Section IV, and Section V contains a conclusion.

### I. The Consistent-Conjectures Hypothesis

The notion of a consistent-conjectures equilibrium is easily explained for a homogeneous-product duopoly in which variable costs are zero and industry demand is linear:  $p = A - B(x_1 + x_2)$ , where  $A > 0$ ,  $B > 0$ ,  $p$  denotes price, and  $x_i$  denotes the output of firm  $i$ . The profit function for firm  $i$  is  $x_i(A - Bx_1 - Bx_2)$ .

The first-order condition for the profit-maximization problem for firm  $i$  is

$$(1) \quad A - Bx_j - 2Bx_i - Bx_i\lambda_j = 0, \\ (i=1,2; j \neq i),$$

where  $\lambda_j \equiv dx_j/dx_i$ . The conjectural variation  $\lambda_j$  is assumed to be a constant.<sup>2</sup>

The two equilibrium outputs cannot be determined from the two equations in (1) unless the  $\lambda_j$  conjectural-variation parameters can be determined. To do this, Bresnahan uses a consistency condition that the actual profit-maximizing reaction of the  $i$ th firm's output to a change in  $x_j$  must be equal to the  $\lambda_i$  conjecture that characterizes the beliefs of firm  $j$ . Suppose that  $x_j$  changes by an amount of  $dx_j$ . Then Bresnahan computes the  $i$ th firm's profit-maximizing re-

sponse to this change by totally differentiating equation (1) to obtain

$$(2) \quad -Bdx_j - 2Bdx_i - B\lambda_j dx_i = 0, \\ (i=1,2; j \neq i).$$

Dividing (2) by  $dx_j$  and using the definition of  $\lambda_i$ , one can express (2) as

$$(3) \quad -B - 2B\lambda_i - B\lambda_j\lambda_i = 0, \\ (i=1,2; j \neq i).$$

It follows from the two equations in (3) that  $\lambda_i = \lambda_j = -1$ . Then (1) implies that  $x_i + x_j = A/B$ , so price and profits are zero for the consistent-conjectures equilibrium in this example.<sup>3</sup> Note that the industry output equals  $A/B$ , but the consistent-conjectures equilibrium outputs need not be equal in this example. This is because the graphs of the reaction functions that satisfy the consistency requirement in the example are overlapping straight lines.

<sup>3</sup>The CCE is a pair of outputs,  $x_1$  and  $x_2$ ; each firm knows the other's output with certainty and each firm calculates that a deviation will be unprofitable given the conjectured reaction of the other. The word conjecture itself connotes uncertainty, so it is natural to consider whether the derivation of the CCE in the text is affected by such uncertainty. Suppose that firm  $i$  is uncertain about the  $j$ th firm's reaction, and let the  $i$ th firm's conjectures about the other firm's reaction be represented by a random variable  $\tilde{\lambda}_j$  with expected value  $\lambda_j$ . Also,  $U_i(\cdot)$  indicates a utility function for firm  $i$ , and  $E_i\{\cdot\}$  indicates an expectation with respect to the  $i$ th firm's subjective distribution for  $\tilde{\lambda}_j$ . The first-order necessary condition for the maximization of expected utility is

$$E_i \left\{ U'_i \left( Ax_i - Bx_i x_j - Bx_i^2 \right) \right. \\ \left. \times \left[ A - Bx_j - 2Bx_i - Bx_i \tilde{\lambda}_j \right] \right\} = 0.$$

Note that the random variable  $\tilde{\lambda}_j$  appears only in the square brackets in the first-order condition. This is because the only uncertainty for firm  $i$  in this analysis is about the  $j$ th firm's reaction to a change in  $x_i$ . It follows from this observation and an assumption that  $U'_i(\cdot) > 0$  that the necessary condition above reduces to equation (1) in the text with  $\lambda_j = E_i\{\tilde{\lambda}_j\}$ .

<sup>2</sup>Bresnahan shows that the consistent conjectural variations will be constants when the profit function is quadratic.

The consistent-conjectures equilibrium concept can be applied when decision variables are prices and there are more than two firms. When the demand is linear, the product is homogeneous, and all firms have identical constant average costs, Morton Kamien and Nancy Schwartz (1983) show that the *CCE* price equals average cost and profits are zero regardless of the number of firms and regardless of whether the decision variables are prices or quantities. The predicted "competitive" result in all cases other than monopoly in this context is the basis of the design of the experiment discussed in Section III.

## II. Evidence from Previous Experiments

The first question that should be addressed is whether the popular static Nash equilibrium approach can explain behavior in multiperiod market experiments. F. Trenery Dolbear et al. (1968) reported data showing that behavior in multiperiod duopoly experiments deviates systematically from a static Nash equilibrium. Their subjects were students who chose prices simultaneously at the beginning of each "period." I will only discuss the "complete information" experiments in which subjects were given a payoff table that relates price choices to payoffs in pennies.<sup>4</sup> The subject's price choice determined a row in the table, and the average of the prices of the subject's competitors determined a column. The payoff entries in the table were computed with a quadratic profit function that resulted from a demand function with some product differentiation. Payoffs were rounded off to the nearest penny, and as a result, there were two symmetric Nash equilibria *in prices* at common prices of 17 or 18. If subjects had been able to collude, they could have maximized their

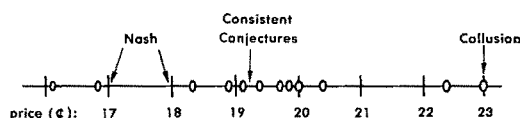


FIGURE 1. PRICE AVERAGES FOR THE 12 DUOPOLY EXPERIMENTS WITH COMPLETE INFORMATION REPORTED IN DOLBEAR ET AL.

joint profit by raising prices to 23. However, subjects were not able to communicate.

In each market experiment, the subjects made simultaneous price decisions 15 times, but they were not told the number of repetitions in advance. The average price for each experiment was obtained by averaging all prices for periods 8 through 12. There were 12 duopoly experiments with complete information, and the average price in each experiment is represented by a large dot on the horizontal price scale in Figure 1. The average price across all 12 experiments was 19.5, and Dolbear et al. concluded that these data indicate some tacit collusion in the sense that average prices and profits exceed the levels determined by a static Nash equilibrium in prices. Using the parameter values for the Dolbear et al. profit function, my earlier paper (1980) calculated the consistent-conjectures equilibrium price to be 19.2 in this context, and this is quite close to the observed price average.

Of course, these experiments were not designed to test the consistent-conjectures equilibrium, and there are several obvious ways in which the experiments do not provide a satisfactory test of this equilibrium concept. First, the subjects were required to make integer-valued price choices, but the *CCE* price was not an integer. Second, there is not much difference between the static Nash and the *CCE* prices. (This problem was even more severe for the oligopoly experiments with four subjects.) Finally, the word "competitors" in the subjects' payoff table may have suggested a particular type of behavior.<sup>5</sup>

<sup>4</sup>Dolbear et al. also considered an "incomplete information" condition. The average level of price choices was approximately the same under each information condition, but there was less dispersion in the incomplete information experiments. Their paper provides an interesting analysis of the effects of information and the number of sellers on the degree of tacit collusion.

<sup>5</sup>Roger Sherman warned me about using suggestive words, but I made the same mistake myself. In one of my pilot experiments, the term "oligopoly game" ap-

Next, consider the previous section's quantity-choice model with a linear market-demand function and a common, constant average cost. The symmetric, static Nash (Cournot) equilibrium when strategies are output *quantities* will result in a price that is greater than average cost and less than the price resulting from perfect collusion. In contrast, the consistent-conjectures equilibrium in this context will result in competitive outputs that drive price down to average cost and profits to zero. Therefore, homogeneous-product oligopoly experiments with quantity-setting subjects and constant average costs may provide a good opportunity to discriminate between the static Nash and the consistent-conjectures theories.

Lawrence Fouraker and Sidney Siegel (1963) reported the results of some complete-information duopoly and triopoly experiments with these characteristics. The columns in their payoff table corresponded to a subject's own output choices, which were integers between 8 and 32. The row was determined by the "Quantity produced by my competition," and this quantity could be between 8 and 64. Outputs between 33 and 64 were actually possible in the triopoly experiments because the "competition" consisted of two other subjects. For a duopoly, the collusive industry output was 30 (15 per subject), the theoretical Nash/Cournot industry output was 40 (20 per subject), and the competitive industry output was 60 (30 per subject). As indicated in the previous section, 60 is the output predicted by the CCE in this context. Fouraker and Siegel do not seem to have noticed that the rounding off of payoffs to the nearest half-penny resulted in two symmetric Nash equilibria: one

at an industry output of 40 and another at an industry output of 44.<sup>6</sup>

There were 16 complete-information duopoly experiments in this series (Experiment 10). Instead of averaging, Fouraker and Siegel used the subjects' decisions in the twenty-first period as an indicator of equilibrium behavior. The period-21 industry outputs were scattered fairly uniformly over the range from the collusive industry output (30) to the competitive (and CCE) industry output (60).<sup>7</sup>

The failure of outputs to rise to the CCE level in many markets may have been due to the fact that the profit was zero because price equaled average cost at this level. Subjects were told in the instructions that if they follow instructions and make "appropriate decisions," they "may earn an appreciable amount of money...but poor choices will result in small or no profit to you." Thus there is a possibility that the wording of the instructions made it less likely that the CCE result with zero profits would be observed. In my own experiments, subjects often appear to be frustrated after periods of very low profits, and such periods are usually followed by large output reductions that raise profits considerably.

There is, for me, a more compelling reason to expect that the outputs of 30 per duopolist would not be frequently observed in the Fouraker-Siegel duopoly experiments. Note that each subject is restricted to choose an output that is less than or equal to 32. The payoff table used by Fouraker and Siegel shows profits for values of the output of the "competition" between 8 and 64. In my opinion, each subject in the duopoly experiments was likely to realize that the outputs from 33 to 64 were irrelevant and, of course, no outputs above 32 were ever observed. If the output of the competition is less than 33, then it is a property of their table that any output below 28 will guarantee the subject a positive profit, regardless of what the competitor does. This truncation of the relevant payoff table caused by exogenous limits on

---

peared on the receipt form to be completed by subjects at the end of the experimental session. This form was passed out at the beginning of the experiment, and one of the subjects who noticed the oligopoly phrase later remarked that the phrase "gave it away." He remembered seeing an assertion in a textbook that oligopolists would collude to maximize joint profit. This subject was in the only duopoly pair (out of four pairs) that was able to reach the collusive output combination in the first market experiment. All data from this pilot experiment were disregarded, and the wording of the receipt form was changed.

<sup>6</sup>See the profit table in their appendix IV.

<sup>7</sup>The industry outputs in period 21 were 25, 30, 30, 32, 33, 38, 39, 40, 40, 44, 45, 49, 50, 55, 59, and 60.

output choices implies that the *CCE* profit of zero can be strictly dominated.<sup>8</sup> In particular, if both subjects were choosing outputs of 30 and earning no profit, then either one could cut output to 15 and earn at least 7.5 cents per period because the other seller's output cannot exceed 32.

This truncation argument does not apply in the triopoly experiments because the competition consists of two subjects, and there is no output decision a subject can make that will ensure a positive profit when each of the other two sellers chooses an output of 32. In fact, behavior in the triopoly experiments did seem to be much more competitive. The static Nash-equilibrium industry output for the triopoly was either 45 or 48.<sup>9</sup> The competitive and *CCE* output was 60, and the actual outputs in period 21 for the 11 triopoly markets were 40, 44, 46, 47, 51, 58, 59, 59, 62, 63, and 70. The median industry output of 58 is quite close to the *CCE* prediction of 60. An industry output of 58 with an approximately symmetric output configuration would result in earnings of only \$.02 per subject per period in 1960 dollars.

### III. An Experimental Test of the Consistent-Conjectures Hypothesis

It follows from the above discussion that an experiment designed to test the consistent-conjectures hypothesis should have the following characteristics: (a) potentially suggestive words such as "competitors" should not appear in the instructions and payoff tables, (b) the *CCE* decisions should be integers, (c) the profit per hour per subject at the *CCE* should be reasonable, and

(d) there should be no decision a subject can make that ensures a profit that will always exceed the *CCE* profit level.

#### A. The Payoff Structure

Bresnahan's original analysis of the *CCE* was for a duopoly in which firms' decision variables are quantities. As indicated in the previous section, models with quantity-setting duopolists are convenient because it is easy to choose market parameters that yield very distinct predictions for the Cournot and the consistent-conjectures equilibria. The experiment reported in this section involved subjects who chose output quantities.<sup>10</sup> Subjects' profits depend on these output choices, and the Profit Table (Table 1) was computed from equation (1) with  $A=12$  and  $B=1/2$ . In addition, \$.45 was added to each of the resulting profit entries. A calculus argument can be used to show that the outputs in a symmetric, collusive equilibrium are 6 per subject and the static Nash/Cournot outputs in a symmetric, collusive equilibrium are 8 per subject. Outputs are integer valued in the experiment, but this does not affect the collusive and Nash equilibria. For example, if both subjects choose outputs of 8, then a unilateral, integer-valued deviation will not increase a subject's profit given the Cournot conjecture. Because of the rounding off of profits to the nearest penny, there are also

<sup>8</sup>This is a serious limitation of the Fouraker-Siegel experiments because the main objective of these experiments seemed to have been to determine the proportions of duopoly pairs which could be best classified as either collusive, Cournot, or competitive. The competitive or "rivalistic" outputs of 30 probably did not have a chance. In a different context, James Murphy (1966) has shown that truncation of the payoff table can have a major effect on experimental results.

<sup>9</sup>The output of 45 was implied by the profit-function parameters, but outputs of 16 for each subject constituted a Nash equilibrium for the payoff table that was used.

<sup>10</sup>In modeling *actual* market situations, the choice of how to model firms' strategies (as output quantities, prices, or something else) is very important. The specification of strategies is not arbitrary. In some markets, firms choose catalogue prices independently and then produce to order. Firms in such markets choose prices independently, but quantities may depend on rivals' prices, so the appropriate decision variable is price. In other markets, key production decisions are made independently in advance, but the price at which a firm's output can be sold depends on other's quantities. Those who use models with quantity-setting firms are presumably considering this type of market. Bresnahan's exposition is for quantity decisions, and it is appropriate to conduct an experimental test of the *CCE* with quantity decisions even if one believes that models with quantity-setting firms are unreasonable for most markets. A useful equilibrium concept should have the property that it yields good predictions for both specifications of decision variables, price and quantity.



TABLE 1—PROFIT TABLE<sup>a</sup>  
Your Output

	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	
Other Seller's Choice	22	41	37	33	27	21	13	5	-5	-15	-27	-39	-53	-67	-83	-99	-117	-135	-155	-175
	21	43	40	36	31	25	18	10	1	-9	-20	-32	-45	-59	-74	-90	-107	-125	-144	-164
	20	45	42	39	34	29	22	15	6	-3	-14	-25	-38	-51	-66	-81	-98	-115	-134	-153
	19	47	45	42	38	33	27	20	12	3	-7	-18	-30	-43	-57	-72	-88	-105	-123	-142
	18	49	47	45	41	37	31	25	17	9	-1	-11	-23	-35	-49	-63	-79	-95	-113	-131
	17	51	50	48	45	41	36	30	23	15	6	-4	-15	-27	-40	-54	-69	-85	-102	-120
	16	53	52	51	48	45	40	35	28	21	12	3	-8	-19	-32	-45	-60	-75	-92	-109
	15	55	55	54	52	49	45	40	34	27	19	10	0	-11	-23	-36	-50	-65	-81	-98
	14	57	57	57	55	53	49	45	39	33	25	17	7	-3	-15	-27	-41	-55	-71	-87
	13	59	60	60	59	57	54	50	45	39	32	24	15	5	-6	-18	-31	-45	-60	-76
	12	61	62	63	62	61	58	55	50	45	38	31	22	13	2	-9	-22	-35	-50	-65
	11	63	65	66	66	65	63	60	56	51	45	38	30	21	11	0	-12	-25	-39	-54
	10	65	67	69	69	69	67	65	61	57	51	45	37	29	19	9	-3	-15	-29	-43
	9	67	70	72	73	73	72	70	67	63	58	52	45	37	28	18	7	-5	-18	-32
	8	69	72	75	76	77	76	75	72	69	64	59	52	45	36	27	16	5	-8	-21
	7	71	75	78	80	81	81	80	78	75	71	66	60	53	45	36	26	15	3	-10
6	73	77	81	83	85	85	85	83	81	77	73	67	61	53	45	35	25	13	1	
5	75	80	84	87	89	90	90	89	87	84	80	75	69	62	54	45	35	24	12	
4	77	82	87	90	93	94	95	94	93	90	87	82	77	70	63	54	45	34	23	

<sup>a</sup>Shown in pennies

two asymmetric Nash equilibrium configurations: one with outputs of 7 and 9 and another with outputs of 6 and 10. In all cases, however, the industry output is 16 in a Nash equilibrium.

It follows from the calculations in Section I that the consistent conjecture is -1 in this context, and any combination of outputs that sum to 24 constitutes a CCE. These output combinations lie on the diagonal with \$.45 profits in the Profit Table. Starting on the diagonal, if one subject increases or decreases output by an integer amount, the other subject is conjectured to make an equal output change in the opposite direction. Thus the new output pair would again be on the \$.45 profit diagonal, so the deviation would not increase the subject's profit, given the consistent conjecture.

The collusive industry output of 12 yields earnings of \$.81 per subject, the static Nash/Cournot industry output of 16 yields earnings of \$.77 per subject in the symmetric case, and the CCE industry output of 24 yields earnings of \$.45 per subject. The experiment was not designed to distinguish noncooperative and collusive behavior, but neither of these modes of behavior yields outputs and profits that are close to those

implied by the consistent-conjectures hypothesis in this context.<sup>11</sup> The high output levels (13 to 22) were included so that no output decision would guarantee a profit that exceeds the CCE level of \$.45 per period.

The \$.45 can be thought of as a normal rate of return when price equals average cost and economic profits are zero. Subjects were also given an initial stake of \$.50 to cover any early losses. The announcement used to solicit subjects stated: "Although earnings cannot be predicted precisely, they will average about \$6 per hour." The experiments were run at a pace of about 13 periods per hour, so the \$.50 stake and the CCE profit of \$.45 per period would result in earnings of about \$6 per hour.

<sup>11</sup>An increase in the *A* parameter will increase the spread between the Cournot and collusive output decisions, but this will increase profits and make the experiments more expensive to run. The use of a fixed cost to lower all profit entries is not possible because the profit at the consistent-conjectures equilibrium should be sufficiently positive. A reduction in the *B* parameter will also increase the spread between the Cournot and collusive outputs, but the resulting flatness in the payoff structure results in multiple Cournot equilibria when profits are rounded off to the nearest penny.

### B. Subjects and Procedures

The subjects were students in introductory and intermediate economics classes at the University of Minnesota. The instructors in these classes had not discussed experimental economics or formal oligopoly theory. The subjects had no previous experience with economics experiments.

Subjects were given about 10 minutes to read the instructions, which are available from the author on request. An additional paragraph in the instructions was read aloud by one of the people conducting the experiments. The purpose of this additional paragraph was to convince the subjects that the "other seller" in "a nearby room" was actually a person (not a computer).

The subjects were also given a Decision Sheet that revealed the "position number" of the other seller in that subject's market. The other sellers were seated in a separate room. First there was a "trial period," in which subjects marked their "output choices" on their Decision Sheets. Then they were told the output choice of the other seller, and they were asked to use the payoff table to compute both their own and the other seller's profit. This allowed us to check the subjects' understanding of the payoff table without suggesting anything by the use of hypothetical outputs to illustrate the computation of profits. In each subsequent period, we collected the Decision Sheets, computed profits, and paid the profits earned before the beginning of the next period. Subjects in the same room were spaced so that they would not be able to see exactly how much money others were earning. Subjects were also invited to write brief "explanations" of their decisions on their Explanation Sheet.

Subjects will naturally be curious about when the experiment will end, and I think the best way to deal with this is to be explicit about the stopping rule. A random stopping rule was used to avoid end effects. Subjects were told that there would be at least 7 periods and that there was a probability of  $1/6$  that period 7 and each following period would be the final period. The final period was determined by a six on the throw of a die, but we used the same sequence of die

throws for all subjects. The throw of the die was recorded on the Decision Sheet.

There were 24 subjects that will be labeled  $S1$ ,  $S2$ , etc. There were 12 initial pairings of subjects, and all subjects participated in a "first market" that was terminated by a throw of the die after 13 periods for all pairs. In order to check for experience effects, 16 of these subjects were rematched and given a new Decision Sheet with the new position number of the other seller. A different sequence of throws of the die was used, and this "second market" was terminated after 9 periods.

### C. The Data

The output choices for the 24 subjects who participated in the first market are shown in Table 2, and choices for the 16 experienced subjects who participated in the second market are shown in Table 3. The final-period industry inputs ( $x_1 + x_2$ ) for all duopoly pairs are plotted along the market demand curve in Figure 2. There was some collusive behavior resulting in outputs of 6 per subject, and there was rivalistic behavior resulting in industry outputs greater than the static Nash/Cournot industry output of 16. Regardless of whether the first-market and second-market data are considered separately or together, the mean and median (or medians) of the final-period industry outputs are between 14 and 16. Earnings averaged about \$8.50 per subject per hour.

The data are clearly inconsistent with the CCE prediction of an industry output of 24, in my opinion. None of the final-period industry outputs exceed 21. There was only one pair of subjects ( $S7$  and  $S2$  in the second market) with combined outputs that were often closer to the CCE level of 24 than to the static Nash/Cournot level of 16. The occasional high outputs of other subjects usually appear to be attempts to punish a rival for not reducing output. For example, subject  $S3$  had been in a collusive duopoly in the first market, but  $S3$  was not able to induce  $S6$  to collude in the second market. Apparently frustrated,  $S3$  increased output from 6 to 19 in period 4 and then returned to 6 in period 5.

TABLE 2—FIRST-MARKET OUTPUT CHOICES FOR SUBJECTS S1–S24<sup>a</sup>

Period	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
1	10	6	10	8	8	10	12	8	10	8	10	10
2	9	10	8	10	8	8	14	9	9	10	4	8
3	10	11	8	6	7	7	13	6	11	9	10	10
4	8	4	6	7	7	9	13	7	9	8	4	8
5	8	10	6	6	8	7	11	8	8	7	10	10
6	10	9	6	6	8	10	11	10	7	7	7	10
7	8	10	6	6	10	8	9	10	7	7	7	8
8	9	8	6	6	10	8	11	9	8	7	7	8
9	10	10	6	6	10	8	10	10	7	10	7	8
10	9	9	6	6	9	9	10	9	8	10	22	8
11	8	8	6	6	9	13	10	9	9	9	7	10
12	7	7	6	6	9	6	10	9	8	8	7	8
13	6	6	6	6	9	8	10	8	7	7	7	8
	S13	S14	S15	S16	S17	S18	S19	S20	S21	S22	S23	S24
1	8	9	7	10	8	7	9	10	6	9	8	4
2	7	8	9	13	6	8	5	9	6	8	9	5
3	6	9	6	7	8	6	9	9	6	8	7	9
4	9	8	7	9	8	9	8	8	6	8	8	13
5	8	8	7	8	8	6	6	9	8	8	11	9
6	7	8	14	6	7	9	8	9	8	8	10	8
7	8	8	8	11	8	6	10	9	8	6	9	7
8	8	8	7	5	8	8	9	8	7	6	8	7
9	8	8	6	6	8	9	7	8	6	7	7	8
10	8	8	6	5	6	10	9	8	6	6	7	8
11	8	8	6	6	8	7	8	8	6	6	8	7
12	8	8	6	6	8	6	8	8	6	6	7	8
13	8	8	6	6	8	8	8	8	6	6	8	6

<sup>a</sup>Subject S1 was paired with S2, S3 with S4, etc.TABLE 3—SECOND-MARKET OUTPUT CHOICES FOR SUBJECTS S1–S16<sup>a</sup>

Period	S1	S4	S3	S6	S5	S8	S7	S2	S9	S12	S11	S14	S13	S16	S15	S10
1	6	6	6	9	9	12	11	8	8	10	7	7	6	10	10	8
2	6	6	6	9	9	12	11	10	9	9	7	7	8	10	8	9
3	6	6	6	9	10	9	11	10	8	9	7	7	8	6	7	8
4	6	6	19	9	9	9	11	10	8	8	7	7	6	8	6	7
5	6	6	6	8	9	8	11	9	7	8	7	7	8	9	6	7
6	6	6	7	8	8	8	11	11	7	8	7	7	8	8	7	7
7	6	6	6	8	8	8	11	10	8	8	7	7	8	8	7	7
8	6	6	8	8	7	8	11	10	8	8	7	7	8	8	7	7
9	6	6	8	10	8	7	11	10	8	8	7	7	8	8	7	7

<sup>a</sup>Subject S1 was paired with S4, S3 with S6, etc.

A statistical analysis should begin with a consideration of why some duopoly pairs are more collusive than others. Variations in market outcomes may be due to variations in variables not included in the oligopoly models discussed above, variables such as individuals' willingness to experiment with output changes. Suppose that individuals'

characteristics are independent drawings from some population of possible characteristics. Then it is natural to think of final-period industry outputs for either the first or second market (not both together) as being independent realizations of a random variable. In the following discussion, the 8 final-period industry outputs in the second mar-

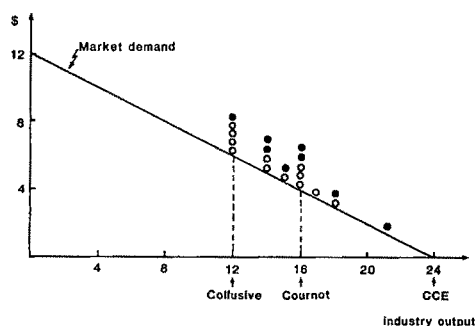


FIGURE 2. FINAL-PERIOD INDUSTRY OUTPUT FOR DUOPOLY MARKETS

Note: ○ First-market duopoly output  
● Second-market duopoly output

ket will be denoted by  $Q_1, Q_2, \dots, Q_8$ , and the vector of these outputs will be denoted by  $\bar{Q}$ . Consider a family of hypotheses of the form:  $Pr\{Q_i < y\} \leq 1/2$  for some  $y > 21$ ;  $i = 1, \dots, 8$ . This family includes a hypothesis that the median of the industry outputs is 24, the theoretical prediction of the consistent conjectures equilibrium. Let  $H_y$  denote a particular hypothesis in this family that corresponds to a particular value of  $y$ . It can be seen from a binomial probability table that  $Pr\{\bar{Q} | H_y\} \leq .0039$  because all 8 industry outputs are less than 21. However, a rejection of  $H_y$  using a classical hypothesis test would be misleading if there were no other hypothesis that is reasonable given the data observed. But there are many reasonable alternatives in this case. For example, consider a hypothesis  $H_{16}$ :  $Pr\{Q_i < 16\} = 1/2$ ,  $i = 1, \dots, 8$ . This hypothesis implies that a median of the distribution is 16, the theoretical prediction of the static Nash equilibrium. It follows from simple binomial probability calculations that  $Pr\{\bar{Q} | H_{16}\} = .2734$ , so the likelihood ratio is greater than  $.2734/.0039$ . If the posterior probabilities for  $H_{16}$  and  $H_y$  are denoted by  $Pr\{H_{16} | \bar{Q}\}$  and  $Pr\{H_y | \bar{Q}\}$ , respectively, then the ratio  $Pr\{H_{16} | \bar{Q}\} / Pr\{H_y | \bar{Q}\}$  is more than 70 times as great as the corresponding ratio of prior probabilities. A Bayesian analysis of the final-period outputs for the first-market experiments yields even stronger conclusions.

#### IV. A Single-Period Duopoly Experiment

The experimental design discussed above induces an infinite horizon in which the probability of termination determines the tradeoff between profit in the current period and profit in the future. In other words, the probability of termination determines the rate of which profits are discounted. If the probability of termination is low enough, subjects may be willing to make unprofitable output reductions in the hope of inducing the other seller to cut output in the future.

Roughly speaking, the behavior in the experiments discussed in Section III can be categorized as either collusive or noncooperative. I expected that an increase in the termination probability from  $1/6$  to 1 would result in no collusion. From a game-theoretic perspective, the static Nash equilibrium is appropriate for single-period games in which subjects are not able to use strategies that are contingent on decisions made in previous periods. Thus, single-period experimental markets would give the static Nash equilibrium its best chance. These markets may also yield even more rivalistic behavior.

I conducted one set of experiments with 12 subjects who participated in a series of 11 single-period duopoly markets with the same payoff table that was used in the multiperiod experiments. The subjects were drawn from a pool of people who had previous experience with a different series of duopoly experiments with different payoff tables. Six subjects were seated in each of two large rooms, and subjects were spaced so that they were unable to determine the "position number" of any other subject in their own room. A research assistant was present in each room at all times. The instructions for these single-period experiments are available from the author on request.

The experiment began with a trial period in which profits were computed but not paid. This was followed by 10 single-period markets. The aggregate data on individual choices for these markets are graphed in Figure 3, and data for particular subjects and their rivals are given in Table 4. The output choices are initially quite diverse, but

TABLE 4—SINGLE-PERIOD EXPERIMENTS: SUBJECTS' OUTPUT CHOICES  
WITH RIVALS' CHOICES SHOWN IN PARENTHESES

Period	S25	S26	S27	S28	S29	S30	S31	S32	S33	S34	S35	S36
Trial	7(8)	8(7)	22(10)	10(22)	5(7)	7(5)	13(7)	7(13)	6(8)	8(6)	10(5)	5(10)
1	5(10)	8(7)	9(6)	10(5)	8(6)	6(9)	11(8)	6(8)	9(7)	8(11)	7(8)	7(9)
2	6(10)	8(9)	9(7)	10(7)	8(8)	10(6)	10(8)	7(9)	9(8)	8(8)	7(10)	8(10)
3	6(7)	8(10)	9(8)	9(9)	8(9)	8(8)	10(8)	7(6)	9(9)	8(9)	8(8)	9(8)
4	6(8)	9(8)	9(9)	8(9)	8(9)	8(9)	9(8)	6(8)	9(8)	8(6)	8(6)	9(9)
5	6(9)	9(9)	9(9)	9(9)	9(9)	8(9)	9(8)	8(9)	9(8)	8(8)	8(8)	9(6)
6	7(9)	8(9)	9(7)	9(8)	9(8)	9(8)	8(9)	8(9)	9(8)	8(9)	8(9)	9(8)
7	7(8)	9(9)	9(9)	9(9)	9(9)	9(9)	9(9)	8(8)	9(9)	8(8)	8(7)	9(9)
8	7(9)	9(8)	9(9)	9(8)	9(7)	8(9)	8(8)	8(9)	9(9)	8(9)	8(8)	9(8)
9	7(8)	9(8)	9(8)	9(9)	9(8)	8(8)	8(7)	8(9)	8(9)	8(8)	8(9)	9(9)
10	7(8)	9(8)	9(8)	8(8)	9(8)	8(10)	8(9)	8(8)	8(7)	8(9)	8(9)	10(8)

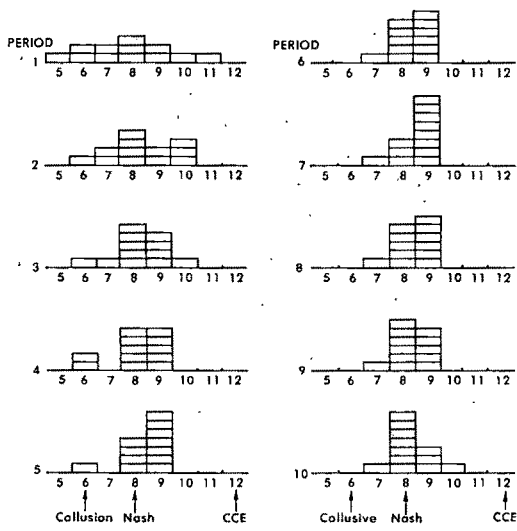


FIGURE 3. FREQUENCY OF INDIVIDUAL OUTPUT CHOICES IN THE SINGLE-PERIOD MARKET EXPERIMENT

by period 7 two-thirds of the subjects are choosing outputs of 9. This is followed by a trend toward the symmetric Nash/Cournot outputs of 8, and 7 of the 12 subjects choose 8 in the final period. As expected, there was no successful collusion in the later periods of this experiment.

The frequency of rivalistic outputs of 9 in the intermediate periods is interesting. First, note that 9 is not very far from a Nash equilibrium in terms of profits. For the range of sellers' outputs in the final periods, any seller with an output of 9 could only increase

profit by \$.01 by switching from 9 to 8. If the outputs are 9 for one seller and 8 for the other, the profit is \$.76 for the high-output seller and \$.73 for the other. At outputs 8 and 8, they each make \$.77. To see why some individuals were willing to give up a penny of profit per period, I looked at the explanation sheets. There were several rivalistic comments about relative profits. For example, one person remarked: "Only a \$.01 loss occurs producing at 9 instead of 8. This keeps the other firm's profits down." This subject did switch to 8 in the final period. Another subject, the only one to have an output of 10 in the final period, remarked in period 4 that when paired "...against a firm with lower output than mine, I make the larger profit, 9 is an interesting number to produce...." However, it is clear that no subject's objective was to maximize the difference between profits; if the other seller produces either 8 or 9, then an output of 12 will maximize the difference between a subject's own profit and that of the other seller. In retrospect, there probably would have been less variability in the data if subjects in these experiments had not been given the complete information necessary to compute the other sellers' profit.

V. Conclusion

In this paper, I compare the theoretical predictions of the consistent-conjectures hypothesis with data for individuals' behavior in several laboratory experiments. In all ex-

periments discussed, subjects simultaneously choose either price or quantity in a sequence of market periods, and subjects are given payoff tables that provide "complete information" about the relationship between decisions and profits for all participants.

My interpretation of the previously published experimental results is—the consistent-conjectures hypothesis provides a good explanation of the price choices made by subjects in the Dolbear et al. experiments, but the predictions of the consistent conjectures and static Nash equilibria are quite close. The predictions of these two equilibria are not close for the Fouraker-Siegel experiments with quantity-setting subjects. The CCE does not provide a good explanation of the output choices in the Fouraker-Siegel duopoly experiments, but its predictions look more reasonable in the triopoly experiments. The poor performance of the CCE in the duopoly case may have been because subjects' profits were zero at the CCE and there were other output choices a duopolist could make that would ensure a strictly positive profit.

This paper reports the results of a new set of duopoly experiments with complete information in which payoffs are positive at the CCE, and there is no decision that can guarantee a profit that exceeds the CCE profit. The consistent-conjectures equilibrium does not provide good predictions in these experiments. The data are more consistent with the Cournot equilibrium, although several duopoly pairs managed to achieve perfect collusion tacitly. Thus, there is at least one simple payoff structure (with homogeneous products, linear demand, and constant average variable costs) in which the CCE predictions are clearly inaccurate.

There are, however, several questions a skeptical reader may wish to consider. First, can laboratory experiments with individual decision makers be used to evaluate theories of the behavior of business firms? Many economists will give a negative answer, but I see nothing in the computation of a consistent-conjectures equilibrium that suggests that the arguments apply to business organizations but not to individuals. One obvious

difference between businessmen and the student subjects is that businessmen have more experience with the markets in which they operate. But when experience has been shown to have a significant impact on behavior in experiments, the effect has been to increase the frequency of collusion.<sup>12</sup> Increased collusion in the experiments reported here would further skew the data away from the "competitive" CCE output prediction.

A second issue is whether the inaccuracy of the CCE prediction derived in Section I is due to something other than the inconsistency of conjectures. In particular, could it be the case that conjectures are consistent but that subjects are maximizing something other than profit? There was a slight tendency toward rivalistic behavior in the single-period experiment, so one may wish to consider an objective function  $R_i$  for the  $i$ th subject of the form:  $R_i = \pi_i + w_i \pi_j$ ; ( $i=1,2$ ;  $j \neq i$ ); where  $\pi_i = x_i(A - Bx_1 - Bx_2)$ ,  $-1 < w_i < 1$ . If the  $w_i$  parameter is zero the subject is a profit maximizer, and as the  $w_i$  parameter approaches  $-1$  the subject becomes very rivalistic and seeks to maximize the difference in profits. The first-order condition analogous to (1) is

$$(4) \quad A - Bx_j - 2Bx_i - Bx_i\lambda_j + w_i[(A - Bx_i - 2Bx_j)\lambda_j - Bx_j] = 0.$$

The consistency condition analogous to (3) is

$$(5) \quad -B - 2B\lambda_i - B\lambda_j\lambda_i + w_i(-B - 2B\lambda_j - B\lambda_j\lambda_i) = 0, \\ (i=1,2; j \neq i).$$

The two equations in (5) imply that  $\lambda_i = \lambda_j$

<sup>12</sup>See Charles Plott (1982) for a discussion of the relationship between experience and collusion in laboratory experiments. Plott also has an excellent summary of the arguments for and against using laboratory experiments to test industrial organization theories.

$= -1$ , so the consistent conjectures are not affected by the possible rivalistic nature of objectives. These conjectures and (4) imply that  $x_i + x_j = A/B$ , so the CCE industry output is unchanged. Thus the inaccuracy of the CCE predictions in this context cannot be attributed to the possibility of nonzero values of the  $w_i$  parameters.

Finally, there is the question of the choice of the rule for ending the experiments. In experiments reported in this paper, the stopping rule was explicit, and a termination probability of  $1/6$  was used in the multi-period experiment. The choice of this particular termination probability was arbitrary because there is no parameter in the theoretical analysis of consistent-conjectures equilibria that corresponds to a termination probability nor is there a discount rate. The CCE concept is not explicitly dynamic; the timing of output deviations, initial reactions, and subsequent reactions by the deviant is not clear. As Perry points out: "The conjectural variation model is a simple static representation of the potentially complex dynamics of an oligopoly, and consistency as defined [in a CCE]...is the simplest adequate static condition for rational behavior in such a model" (p. 200).

The CCE did not provide a satisfactory representation of the dynamics in experimental markets with a termination probability of  $1/6$ . I would expect to observe more collusion and less rivalistic behavior if the termination probability were even less than  $1/6$ . For termination probabilities that exceed  $1/6$ , I would expect behavior to conform more closely to the predictions of the static Cournot model. In the single-period market experiments with a termination probability of 1, the Nash/Cournot equilibrium provided accurate predictions, and there was no tendency to collude.

## REFERENCES

- Bresnahan, Timothy F., "Duopoly Models with Consistent Conjectures," *American Economic Review*, December 1981, 71, 934-45.
- Cyert, Richard M. and DeGroot, Morris H., "Multi-period Decision Models with Alternating Choice as a Solution to the Duopoly Problem," *Quarterly Journal of Economics*, August 1970, 84, 410-29.
- Dolbear et al., F. Trenery, "Collusion in Oligopoly: An Experiment on the Effect of Numbers and Information," *Quarterly Journal of Economics*, May 1968, 82, 240-59.
- Fouraker, L. E., and Siegel, S., *Bargaining Behavior*, New York: McGraw-Hill, 1963.
- Friedman, James W., *Oligopoly and the Theory of Games*, Amsterdam; New York: North-Holland, 1977.
- Green, Edward J. and Porter, R. H., "Noncooperative Collusion Under Imperfect Price Information," *Econometrica*, January 1984, 52, 87-100.
- Holt, Charles A., "Equilibrium Models of Tacit Collusion in Oligopoly Experiments with Price-Setting Firms," Discussion Paper No. 80-138, Center for Economic Research, University of Minnesota, October 1980.
- Kamien, Morton I. and Schwartz, Nancy L., "Conjectural Variations," *Canadian Journal of Economics*, May 1983, 16, 191-211.
- Murphy, James L., "Effects of the Threat of Losses on Duopoly Bargaining," *Quarterly Journal of Economics*, May 1966, 80, 296-313.
- Perry, Martin K., "Oligopoly and Consistent Conjectural Variations," *Bell Journal of Economics*, Spring 1982, 13, 197-205.
- Plott, Charles R., "Industrial Organization Theory and Experimental Economics," *Journal of Economic Literature*, December 1982, 20, 1485-527.

# Rational Expectations and the Limits of Rationality: An Analysis of Heterogeneity

By JOHN HALTIWANGER AND MICHAEL WALDMAN\*

A recurring controversy in economic thought concerns the conflict between the assumption of rationality and the fact that economic agents have limited capacities to process information.<sup>1</sup> For example, this conflict was a factor in the marginalist debate of the 1940's, and is a factor in the challenge to standard economic theory of Herbert Simon and his followers.<sup>2</sup> Recent refinements to the concept of rationality have brought this conflict into sharper focus. That is, rationality no longer simply implies that behavior is determined by the maximization of a well-ordered function. Rather, it now typically implies the expected utility hypothesis of behavior under uncertainty, and the rational expectations hypothesis for the formation of expectations.

Those individuals who have attempted to provide alternatives to the rationality assumption have in general paid close atten-

tion to descriptive realism. In other words, the alternative models of the decision-making process which have been provided tend to match, at least to a first approximation, the manner in which decisions are reached in the real world. One aspect of the real world, however, has generally been ignored in these alternative models. Specifically, these models, as well as most models that contain the rationality assumption, ignore the idea that agents tend to be heterogeneous in terms of information-processing abilities. That is, some agents in our economy are able to process information in a very sophisticated manner, while others are much more limited in their capabilities. In the present paper we attempt to investigate the ramifications of this type of heterogeneity. We do this by analyzing two simple models wherein agents differ in terms of their ability to form expectations. To keep the analysis tractable, in each of the models it is assumed that only two types of agents exist. Agents of the first type have unlimited abilities to form expectations, and thus have correct or rational expectations. These agents will be referred to as "sophisticated." Agents of the second type are limited in their ability to form expectations, and thus have incorrect expectations. These agents will be referred to as "naive." Our goal is to characterize equilibria under this type of heterogeneity, and in the process identify situations in which sophisticated agents have a disproportionately large effect on equilibrium, and situations in which naive agents have a disproportionately large effect.<sup>3,4</sup>

\*Department of Economics, University of California, Los Angeles, CA 90024. We thank Robert Clower, Ian Novos, Mark Plant, Earl Thompson, Susan Woodward, the participants of the Theory Workshop at UCLA, and two anonymous referees for helpful comments, and the Foundation for Research in Economics and Education for financial support. We accept responsibility for any remaining errors.

<sup>1</sup>This conflict does not mean that agents who have limited capacities to process information are not rational under some broad definition of the term, but rather that they are not rational under the literature's definition of the term "rationality." Note, this limited ability to process information is sometimes referred to as bounded rationality. However, because of its frequent association with the related concept termed satisficing, we will refrain from using the term bounded rationality in this paper.

<sup>2</sup>References to the marginalist debate of the 1940's include Robert Hall and Charles Hitch (1939) and Fritz Machlup (1946). References to the work of Simon and his followers include Simon (1959), Richard Cyert and James March (1963), Oliver Williamson (1975), and Richard Nelson and Sidney Winter (1982). See also Roy Radner (1975) for explicit modeling of Simon's ideas.

<sup>3</sup>The term equilibrium in this paper simply refers to the outcome of the model, given an exogenous specification for the expectations of the naive.

<sup>4</sup>Three recent papers which do consider the type of heterogeneity we consider are John Conlisk (1980), Thomas Russell and Richard Thaler (1985), and George



One might question our approach of looking at the equilibria of models that contain agents who have incorrect expectations. Some might argue that, since models with learning frequently converge to rational expectations equilibria, in the types of models we are investigating it only makes sense to investigate rational expectations equilibria.<sup>5</sup> We disagree with this argument for three reasons. First, there are many important economic situations that are not repeated for the agents involved, and for these situations the fact that learning models converge to rational expectations equilibria is irrelevant. Examples include the career choice problem faced by the young, and the related decision concerning whether or not to attend college. Second, there are many economic situations that do repeat; however, for each repetition there is a proportion of agents who have no previous experience with the situation. An example of this is the problem consumers face when they must choose among different computer hardware systems. That is, for each new generation of computers there is likely to be a proportion of buyers who have not previously bought a computer system. Third and finally, empirical evidence does not support the practice of restricting attention solely to rational expectations equilibria (see Kenneth Arrow, 1982, for a survey of this evidence).

In Section I, each agent in the economy is faced with the problem of choosing a single path from a set of two, where the paths exhibit congestion effects. By congestion effects we mean that for any agent  $i$ , the higher is the number of other agents who choose the same path as the one chosen by agent  $i$ , the less well off is agent  $i$ . The model

of this section might best be thought of as a stylized model of either the problem of career choice, or the problem of choosing a road with which to reach some final destination. In this setting we find that sophisticated agents have a disproportionately large effect on equilibrium. That is, congestion effects cause the equilibrium to more closely resemble what occurs when all agents have rational expectations than would be suggested by the relative number of sophisticated agents and naive agents in the population.

The intuition behind this result is as follows. Because of incorrect expectations, the allocation of naive agents to paths is biased relative to how agents are allocated when there are no naive agents. Sophisticated agents, on the other hand, anticipate this behavior and, because of congestion effects, compensate by having their behavior being biased in an exactly opposite manner. Or overall, sophisticated agents have a disproportionately large effect on equilibrium because sophisticated agents anticipate the bias of the naive, and compensate in a way which tends to nullify this bias.

In Section II we again look at the model developed in Section I, but we change the specification so that the model exhibits the reverse of what we previously referred to as congestion effects. In other words, for any agent  $i$ , the higher is the number of other agents who choose the same path as the one chosen by agent  $i$ , the *better* off is agent  $i$ . We will say that situations which exhibit this type of property exhibit synergistic effects. Contrary to what we found in Section I, here it is the naive agents who have a disproportionately large effect on equilibrium. That is, synergistic effects cause the equilibrium to more closely resemble what occurs when all agents are naive than would be suggested by the relative number of sophisticated agents and naive agents in the population.

The intuition behind this result is similar to the intuition given above. Because of incorrect expectations, the allocation of naive agents to paths is again biased relative to how agents are allocated when there are no naive agents. Furthermore, sophisticated agents again anticipate this behavior, but because there are synergistic effects, they now

---

Akerlof and Janet Yellen (1985). The main difference between our paper and these papers is that we focus on the conditions that lead to sophisticated agents having a disproportionately large effect on equilibrium, and also on the conditions that lead to the naive having a disproportionately large effect—an issue not addressed in the above papers.

<sup>5</sup> Examples of papers wherein learning causes convergence to rational expectations equilibria include Cyert and Morris DeGroot (1974), and Stephen DeCanio (1979).

compensate by having their behavior being biased in a manner *similar* to the bias of the naive. Or overall, naive agents have a disproportionately large effect on equilibrium because sophisticated agents anticipate the bias of the naive, and compensate in a way which tends to reinforce this bias.

In Section III we consider a simple model of the chain-store paradox (see Richard Selten, 1978; David Kreps and Robert Wilson, 1982; and Paul Milgrom and John Roberts, 1982, for papers on this issue). The key property of this model is that it is one where reputation effects are potentially important. The basic result that emerges from the analysis of Section III is that, when reputation effects are potentially important, it is possible for either type of agent to be dominant. That is, under some parameterizations sophisticated agents have a disproportionately large effect on equilibrium, while for other parameterizations the naive agents are the ones who have the disproportionately large effect.

One might ask what conclusions can be drawn from the above results with regard to the common practice of assuming that all agents in the economy have rational expectations. Our feeling is that, since agents in the real world are obviously heterogeneous in terms of information-processing abilities, the practice of assuming rational expectations is relatively more defensible when agents who can process information in a very sophisticated manner have a disproportionately large effect on equilibrium. Thus, our results suggest that for the analysis of situations that exhibit congestion effects, there are relatively strong justifications for assuming rational expectations. However, for the analysis of situations that exhibit either synergistic effects or the possibility for reputation formation, the practice of assuming rational expectations would seem to be less defensible.

### I. Choosing Paths with Congestion Effects

Here we consider the problem of agents choosing between two paths, where the paths exhibit congestion effects. By congestion effects we mean that for any agent  $i$ , the higher is the number of other agents who

choose the same path as the one chosen by agent  $i$ , the worse off is agent  $i$ . Examples of real world choice situations that exhibit congestion effects are the career choice problem faced by the young, the related decision concerning whether or not to attend college, and the commuter's problem of choosing between alternate routes.

As indicated above, each agent must choose between two paths, which we denote as  $A$  and  $B$ . Furthermore, it is assumed that this choice of paths is an irreversible choice, and is made simultaneously by all the agents in the population. If agent  $i$  takes path  $A$ , then his utility equals  $r_A(N_A, \theta_A) - c_i$ . Similarly, if agent  $i$  takes path  $B$ , then his utility equals  $r_B(N_B, \theta_B) - (C - c_i)$ . The term  $N_A(N_B)$  denotes the total number of agents who choose path  $A(B)$ , while  $\theta_A$  and  $\theta_B$  are random variables, the realizations of which are unknown at the date the participation decision is made. It is further assumed that agents are expected utility maximizers, which means that we need only be concerned with  $f_A(N_A)$  and  $f_B(N_B)$ , where  $f_A(N_A) = E_{\theta_A}[r_A(N_A, \theta_A)]$  and  $f_B(N_B) = E_{\theta_B}[r_B(N_B, \theta_B)]$ . Formally, assuming that the paths exhibit congestion effects means assuming  $f'_A \leq 0$  and  $f'_B \leq 0$ , where one of these inequalities is everywhere strictly negative. Note,  $c_i$  and  $(C - c_i)$  can be interpreted as representing either agent  $i$ 's underlying preferences for the two paths, or the costs incurred by agent  $i$  in taking each path. This latter interpretation matches well with our description of the model as a model of commuting. That is, when thought of as a model of commuting,  $c_i$  can be interpreted as the distance between agent  $i$ 's housing location and path  $A$ 's entry ramp, while  $(C - c_i)$  can be interpreted as the distance between the housing location and path  $B$ 's entry ramp.

The population consists of a continuum of agents who vary in terms of their values for  $c_i$ . In particular, the distribution of  $c_i$ 's in the population is described by a density  $g(c_i)dc_i$  defined on the interval  $[0, C]$ , where  $g(\cdot)$  is continuously differentiable and non-zero everywhere in the specified interval. There are two types of agents in this population: sophisticated and naive. It is assumed that a proportion  $p$  of the total population is

sophisticated,  $0 \leq p \leq 1$ , while a proportion  $(1-p)$  is naive. Formally, this translates into assuming that the distribution of  $c_i$ 's in the population of sophisticated (naive) agents is described by a density  $h(c_i)dc_i$  ( $j(c_i)dc_i$ ) defined on the interval  $[0, C]$ , where  $h(\cdot) = pg(\cdot)$  ( $j(\cdot) = (1-p)g(\cdot)$ ) everywhere in this interval. Note, we also assume  $f_A(0) > f_B(\int_0^C g(c_i)dc_i) - C$  and  $f_B(0) > f_A(\int_0^C g(c_i)dc_i) - C$ . This pair of assumptions guarantees that, as long as  $0 < p \leq 1$ , the equilibrium will be characterized by some agents taking each path.

The only aspect of the model that remains to be specified concerns expectations. Expectations are relevant in that, prior to deciding which path to take, each agent forms expectations concerning the resulting value for  $f_A(N_A) - f_B(N_B)$ . Sophisticated agents are assumed to have unlimited abilities to form expectations, and thus have correct or rational expectations concerning the resulting value for  $f_A(N_A) - f_B(N_B)$ . On the other hand, naive agents are limited in their ability to form expectations, but all have the same incorrect expectation concerning  $f_A(N_A) - f_B(N_B)$ , where  $V$  denotes this expectation.<sup>6</sup>

The above assumption that all naive agents have the same incorrect expectation is basically a simplifying assumption. There is, however, a reason for assuming that all naive agents have similar biases in their expectations. That is, in the studies of cognitive psychologists it is frequently the case that the expectational errors of agents are correlated across individuals (see Daniel Kahneman, Paul Slovic, and Amos Tversky, 1982).

The rest of this section consists of an analysis of the above model.<sup>7</sup> Let  $D = (f_A(N_A) - f_B(N_B) + C)/2$  and  $D^n = (V + C)/2$ . Given that sophisticated agents have rational expectations, sophisticated agent  $i$

will choose path  $A(B)$  when<sup>8</sup>

$$(1) \quad c_i < (>) D.$$

On the other hand, given the expectations of the naive, naive agent  $i$  will choose path  $A(B)$  when

$$(2) \quad c_i < (>) D^n.$$

Equations (1) and (2) in turn yield<sup>9</sup>

$$(3) \quad N_A = \int_0^D pg(c_i)dc_i + \int_0^{D^n} (1-p)g(c_i)dc_i$$

$$(4) \quad N_B = \int_D^C pg(c_i)dc_i + \int_{D^n}^C (1-p)g(c_i)dc_i$$

Let  $N_A^s(N_B^s)$  denote the number of agents who choose path  $A(B)$  when all are sophisticated (i.e.,  $p=1$ ),  $N_A^n(N_B^n)$  denote the number of agents who choose path  $A(B)$  when all are naive (i.e.,  $p=0$ ), and  $D^s$  denote the value for  $D$  when  $p=1$ . Note, we refer to the equilibrium when  $p=1$  as the pure rational expectations equilibrium, and the equilibrium when  $p=0$  as the pure limited rationality equilibrium. We now proceed to our first proposition (note: all proofs are relegated to the Appendix).

**PROPOSITION 1:** *If  $0 < p < 1$  and  $N_A^s \geq N_A^n$ , then  $N_A \geq pN_A^s + (1-p)N_A^n$ . Note, this also implies that if  $0 < p < 1$  and  $N_B^s \geq N_B^n$ , then  $N_B \geq pN_B^s + (1-p)N_B^n$ .*

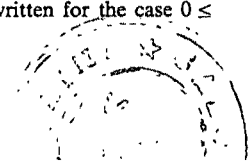
The interpretation of Proposition 1 is straightforward. When agents are heterogeneous in terms of information processing abilities (i.e.,  $0 < p < 1$ ), sophisticated agents have a disproportionately large effect on the number of agents who take each path. That is, in terms of the number of agents who take

<sup>6</sup> Given that agents are risk neutral,  $V$  can be interpreted as the mean of a distribution that describes the naive agents' expectations.

<sup>7</sup> Our analysis derives properties which an equilibrium for this model must display. In a mathematical supplement available from the authors upon request, we demonstrate that, for the specifications of both Section I and Section II, an equilibrium exists and is unique.

<sup>8</sup> Because it concerns a set of agents whose weight is zero, we need not specify what happens when  $c_i = D$ .

<sup>9</sup> Equations (3) and (4) are written for the case  $0 \leq D^n \leq C$ .



each path, the equilibrium more closely resembles the pure rational expectations equilibrium than would be suggested by the relative number of sophisticated agents and naive agents in the population. The next step of the analysis is to consider the social welfare aspects of our model.

Social welfare, denoted  $W$ , is here defined as the sum of the utilities of all the agents in the population. Furthermore, to stay consistent with previous notation,  $W^s$  will denote social welfare when  $p=1$  and  $W^n$  will denote social welfare when  $p=0$ . In terms of social welfare, we feel there are two questions which deserve to be addressed. The first question is analogous to the question addressed in Proposition 1. That is, how does  $W$  compare with  $\bar{W}$ , where  $\bar{W} = pW^s + (1-p)W^n$ ? The second question is, how do the reactions of the sophisticated agents, in response to the behavior of the naive, affect social welfare? That is, do the reactions of the sophisticated tend to drive social welfare towards  $W^s$  or towards  $W^n$ ? To answer this second question we compare  $W$  with  $\hat{W}$ , where  $\hat{W}$  is the social welfare that results when sophisticated agents behave as if  $D$  equals  $D^s$ , while naive agents behave as before. In other words,  $\hat{W}$  is the social welfare that results when, even though agents are heterogeneous, sophisticated agents behave as if all agents were sophisticated.

Before proceeding to our social welfare proposition, it is necessary to define some terms concerning the optimal allocation of agents to paths. Note, because congestion effects create externalities, the equilibrium when all agents are sophisticated does not in general yield an optimal allocation of agents to paths. Suppose there was a social welfare-maximizing government that was all knowing, and that could direct the actions of all the agents in the population. We denote as  $N_A^0(N_B^0)$  the total number of agents such a government would send to path  $A(B)$ .

We can now make comparisons between  $W$  and our two social welfare benchmarks.

**PROPOSITION 2:** Suppose  $f_A'' \leq 0$ ,  $f_B'' \leq 0$ ,  $0 < p < 1$  and that either  $N_A^s, N_A^n \leq N_A^0$  or  $N_A^s, N_A^n \geq N_A^0$ . If  $W^s > W^n$ , then  $W > \bar{W}$  and  $W > \hat{W}$ . If  $W^s < W^n$ , then  $W < \bar{W}$ .

Proposition 2 provides support for the notion that, even in terms of social welfare, sophisticated agents tend to be disproportionately important. That is, given a few restrictions, the following two statements are necessarily true. First, in response to the behavior of the naive, the sophisticated react in a manner which drives social welfare towards  $W^s$  and away from  $W^n$ . Second, for the case  $W^s > W^n$ , it is also true that  $W$  is closer to  $W^s$  than would be suggested by the relative number of sophisticated agents and naive agents in the population. The reason our results are stronger for the case  $W^s > W^n$  is the following. As indicated above, given our restrictions we are able to demonstrate that the response of the sophisticated to the behavior of the naive always drives social welfare towards  $W^s$  and away from  $W^n$ . It is also true that, given our restrictions,  $\hat{W}$  is never less than  $\bar{W}$ . Thus, when  $W^s > W^n$ ,  $W > \hat{W}$  implies  $W > \bar{W}$ . However, when  $W^s < W^n$  the same logic does not apply.

A word is in order concerning our restrictions. First, we assume there are increasing marginal congestion effects, that is,  $f_A'' \leq 0$  and  $f_B'' \leq 0$ . This assumption basically serves to guarantee that social welfare is a single-peaked function of the number of sophisticated agents who take each path—a property which is critical for our proof.<sup>10</sup> Second, we assume that either  $N_A^s, N_A^n \leq N_A^0$  or  $N_A^s, N_A^n \geq N_A^0$ . The role of this assumption is as follows. Our proof relies on the idea that, when the behavior of the naive drives social welfare down, the sophisticated respond by shifting towards first-best optimal behavior—thus driving social welfare up. On the other hand, when the behavior of the naive drives social welfare up, the sophisticated respond by shifting away from first-best optimal behavior—thus driving social welfare down. The role of the restriction that either  $N_A^s, N_A^n \leq N_A^0$  or  $N_A^s, N_A^n \geq N_A^0$  is primarily to guarantee that the response of the sophisticated, relative to first-best optimal behavior,

<sup>10</sup> There is evidence, however, that suggests that, at least for our transportation interpretation, it is reasonable to assume increasing marginal congestion effects (see Theodore Keeler and Kenneth Small, 1977).

is consistent with the above. For example, suppose the restriction was violated and the naive behaved in a manner that tended to lower social welfare. Even though the sophisticated would respond by having behavior being biased in an exactly opposite manner, given that our restriction is violated, this could actually entail having the behavior of the sophisticated move away from first-best optimal behavior.<sup>11</sup>

One interesting implication of Proposition 2 concerns the following issue. Consider a choice situation where a proportion of the population is naive, and the naive behave in a manner that lowers social welfare. Proposition 2 tells us that, if the environment exhibits congestion effects, then there may be little reason to be concerned about the incorrect expectations of the naive. That is because, in response to the behavior of the naive, the sophisticated tend to react in a manner which reduces the social welfare loss attributable to these incorrect expectations.

Finally, we would like to point out that in addition to providing insights concerning the interaction of naive and sophisticated agents, this model can also be used to provide insights concerning the interaction of informed and uninformed agents (see Sanford Grossman and Joseph Stiglitz, 1980, and Russell Cooper and Thomas Ross, 1984). That is, the model can be reinterpreted as one where all agents have rational expectations, but only some agents are informed about a relevant piece of information. For example, consider

the problem of commuters having to choose which of two routes to travel, where only some agents are aware that one of the routes has a lane blocked off. The model analyzed in this section can be reinterpreted as a model of just this problem. Under this interpretation the sophisticated agents become the agents who are informed of the closed lane, while the naive agents become the uninformed agents. The conclusion that can be drawn concerning the interaction of informed and uninformed agents is basically the same conclusion as drawn above for sophisticated and naive agents. That is, when there are congestion effects, informed agents tend to have a disproportionately large effect on equilibrium.

## II. Choosing Paths with Synergistic Effects

Here we consider the problem of agents choosing between two paths, where the paths exhibit synergistic effects. By synergistic effects we mean that for any agent  $i$ , the higher is the number of other agents who choose the same path as the one chosen by agent  $i$ , the better off is agent  $i$ . An example of a real world choice situation that exhibits synergistic effects is the problem faced by consumers in choosing a computer hardware system. This choice problem exhibits synergistic effects in that the larger the number of individuals who purchase a particular system, the greater will be the subsequent availability of computer peripherals and software for that system. Other examples include the problems faced by consumers in choosing between a cassette and an 8-track tape player, and the choice between the Beta and VHS formats for video cassette recorders. Similar to the computer example, these choice problems exhibit synergistic effects because the larger the number of individuals who purchase a particular system, the greater will be the subsequent variety of tapes available for that system.

To investigate the problem of agents choosing between paths when synergistic effects are present, we analyze a variant of the model developed in the previous section. The new assumptions are as follows. First we assume  $f'_A \geq 0$ ,  $f'_B \geq 0$  and that one of the

<sup>11</sup>An additional reason for imposing the restriction that either  $N_A^s, N_A^n \leq N_A^O$  or  $N_A^s, N_A^n \geq N_A^O$  is that without the restriction some of the comparisons we make in Proposition 2 can be quite misleading. Consider, for example, the following discussion concerning  $\hat{W}$ . When the restriction that either  $N_A^s, N_A^n \leq N_A^O$  or  $N_A^s, N_A^n \geq N_A^O$  is satisfied,  $\hat{W}$  will necessarily lie strictly between  $W^s$  and  $W^n$ . Thus, if the restriction is satisfied, our comparison will suggest that the sophisticated are disproportionately important whenever the response of the sophisticated moves social welfare towards  $W^s$ . When the restriction is not satisfied, however,  $\hat{W}$  can actually lie outside the interval defined by  $W^n$  and  $W^s$ . Thus, if the restriction were not satisfied, our comparison could actually suggest that the naive are disproportionately important even when it is the case that  $W = W^s$ .

two is everywhere strictly positive. Second,  $[f'_A(\int_0^z g(c_i)dc_i) + f'_B(\int_z^C g(c_i)dc_i)]g(z)/2 < 1$  for all  $0 \leq z \leq C$  is assumed. The first assumption insures that the model now exhibits synergistic effects; the second assumption eliminates the possibility of multiple equilibria (note: in Section I no similar assumption was needed to rule out the possibility of multiple equilibria).

Under this specification equations (1)–(4) of the previous section continue to hold. We now proceed to the Section II analogue of Proposition 1.

**PROPOSITION 3:** *If  $0 < p < 1$  and  $N_A^n \geq N_A^s$ , then  $N_A = pN_A^s + (1-p)N_A^n$ . Note, this also implies that if  $0 < p < 1$  and  $N_B^n \geq N_B^s$ , then  $N_B = pN_B^s + (1-p)N_B^n$ .*

The interpretation of Proposition 3 is straightforward. When agents choose among paths which exhibit synergistic effects, naive agents, rather than sophisticated agents, have a disproportionately large effect on the number of agents who take each path. That is, in terms of the number of agents who take each path, the equilibrium more closely resembles the pure limited rationality equilibrium than would be suggested by the relative number of sophisticated agents and naive agents in the population.

We feel this result is especially interesting in that it provides a rationale for the persistent dominance of IBM in the computer industry. As mentioned earlier, one example of a choice situation that exhibits synergistic effects is the problem of choosing a computer hardware system. This problem exhibits synergistic effects because the availability of computer peripherals and software for a system depends on the number of individuals who purchase the system. Now, suppose each time a new generation of computers comes onto the market, a segment of the population anticipates that the market shares of the different brands will be the same as for the previous generation of computers. Because of the synergistic properties of the market, in such a situation brands that are successful in one period would have a high probability of having their success repeated. There are two reasons for this. First, consumers who extrapolate will tend to

purchase brands with previous high market shares. Second, given Proposition 3, it is also true that more sophisticated consumers will tend to behave in this manner. Thus, the persistent dominance of IBM in the computer industry may very well be due to the synergistic properties of the market.

The next step of the analysis is to consider again the social welfare aspects of the model. In what ensues  $W$ ,  $\bar{W}$ ,  $\hat{W}$ ,  $N_A^O$  and  $N_B^O$  will be defined as in Section I. For the analysis of synergistic effects, however, it is also necessary to define some terms concerning the following second-best problem. Suppose there was a social welfare-maximizing government which was all knowing, but which could only direct the actions of the sophisticated agents in the population. We denote as  $N_A^O(N_B^O)$  the total number of agents who would wind up at path  $A(B)$ , given a government with this limited ability to direct behavior.

We can now make comparisons between  $W$  and our two social welfare benchmarks.

**PROPOSITION 4:** *Suppose  $f_A'' \leq 0$ ,  $f_B'' \leq 0$ ,  $0 < p < 1$  and that either  $N_A^s, N_A^n \leq N_A^O$  and  $\hat{N}_A \leq N_A^O$ , or  $N_A^s, N_A^n \geq N_A^O$  and  $\hat{N}_A \geq N_A^O$ , where  $\hat{N}_A = pN_A^s + (1-p)N_A^n$ . If  $W^n < W^s$ , then  $W < \bar{W}$  and  $W < \hat{W}$ . If  $W^n > W^s$ , then  $W > \bar{W}$ .*

Proposition 4 provides support for the notion that, even in terms of social welfare, naive agents tend to be disproportionately important. That is, given a few restrictions, the following two statements are necessarily true. First, in response to the behavior of the naive, the sophisticated react in a manner which drives social welfare towards  $W^n$  and away from  $W^s$ . Second, if  $W^n > W^s$ , then  $W$  is closer to  $W^n$  than would be suggested by the relative number of sophisticated agents and naive agents in the population. The reason our results are stronger for the case  $W^n < W^s$  is similar to the reason our results in Proposition 2 were stronger for the case  $W^n < W^s$ . The only difference is that now  $\bar{W}$  is never less than  $\hat{W}$ .

A word is in order concerning our restrictions. First, we assume there are decreasing marginal synergistic effects, that is,  $f_A'' \leq 0$  and  $f_B'' \leq 0$ . As was the role of the assump-

tion of increasing marginal congestion effects in Proposition 2, this assumption guarantees that social welfare is a single-peaked function of the number of sophisticated agents who take each path. Second, we assume that either  $N_A^s, N_A^n \leq N_A^o$  and  $\hat{N}_A \leq N_A^o$ , or  $N_A^s, N_A^n \geq N_A^o$  and  $\hat{N}_A \geq N_A^o$ . This assumption serves the same role as the assumption in Proposition 2 that either  $N_A^s, N_A^n \leq N_A^o$  or  $N_A^s, N_A^n \geq N_A^o$ . One question that arises is, why is the condition more restrictive in Proposition 3? The answer is that, when there are congestion effects (assuming that  $N_A^s, N_A^n \leq (\geq) N_A^o$  guarantees that  $\hat{N}_A \leq (\geq) N_A^o$ ). A second question is, what is the role played by the additional restriction? Propositions 2 and 3 work because, under our restrictions, the following statement is satisfied. When the presence of naive agents causes the behavior of the sophisticated to move away from (towards) first-best optimal behavior, social welfare tends to be driven down (up) or below (above) our social welfare benchmarks. The role of our additional restriction is simply to guarantee that the above statement is valid. That is, if the restriction is violated, then we have the possibility of a standard second-best problem wherein having the sophisticated move away from (towards) first-best optimal behavior might actually increase (decrease) social welfare.

As in the previous section, we can consider what our analysis implies for a choice situation where a proportion of the population is naive, and the naive behave in a manner that tends to lower social welfare. Proposition 4 tells us that, if such a choice environment exhibits synergistic effects, then there may be reason to be quite concerned about the incorrect expectations of the naive. That is because, in response to the behavior of the naive, the sophisticated now react in a manner that tends to increase the social welfare loss attributable to these incorrect expectations.

Finally, in concluding the previous section we noted that in addition to providing insights concerning the interaction of naive and sophisticated agents, the model of Sections I and II can also be used to provide insights concerning the interaction of informed and uninformed agents. That is, the model can be reinterpreted as one where all

agents have rational expectations, but only some agents are informed of a relevant piece of information. Our conclusion regarding this point in Section I was that, when there are congestion effects, informed agents tend to have a disproportionately large effect on equilibrium. In contrast, the conclusion we can draw from the analysis of Section II is that, when there are synergistic effects, the uninformed agents tend to be the ones who are disproportionately important.

### III. The Chain-Store Paradox Revisited

Here we analyze a variant of a game initially studied by Selten. The game is basically described as follows. Let there be a monopolist who operates in  $T$  separate markets, and who sequentially faces a different potential entrant in each market. Furthermore, in each market let the potential entrant move first by deciding whether or not to enter the market. If the potential entrant decides not to enter, then there are no further moves by players in that market. If entry does occur, however, then the monopolist has to decide whether to cooperate or to act aggressively. Finally, let payoffs satisfy the following three conditions. First, if entry has occurred in market  $k$ , then both the monopolist and the entrant receive higher profits in market  $k$  if the monopolist acts cooperatively. Second, for each market  $k$ , the monopolist receives even higher profits if entry does not occur. Third, for each market  $k$ , if the monopolist is sure to act aggressively then the potential entrant is better off not entering.

Selten referred to this game as the chain-store game, and pointed out that the game embodies a paradox. On the one hand, intuition suggests that if such a game were to actually be played, the monopolist would likely act aggressively whenever entry occurs in an early market, and in turn this would limit the number of markets in which entry occurs. On the other hand, game theory predicts that entry will occur in every market, and that the monopolist will cooperate each time entry occurs.

Recently, Kreps-Wilson and Milgrom-Roberts have demonstrated one way in which the paradox can be resolved. Their basic idea

is that, if there is even a small probability that the monopolist always acts aggressively, then the resulting equilibrium closely resembles the intuitive equilibrium specified above. That is, the monopolist always acts aggressively if entry occurs in an early market, and this in turn deters entry in early markets.<sup>12</sup>

In this section we investigate what happens when the decision-making process of the potential entrants is varied, rather than the above approach of varying the decision-making process of the monopolist. Specifically, we investigate how equilibrium is affected when, for each market, there is some probability that the potential entrant is naive and forms expectations by simply extrapolating from previous behavior. The basic result that emerges from our analysis is that, when reputation effects are potentially important, it is possible for either type of agent to be dominant. That is, under some parameterizations, sophisticated agents will have a disproportionately large effect on equilibrium, while for other parameterizations, the naive agents will be the ones who are disproportionately important.

As stated previously, we consider a monopolist who operates in  $T$  separate markets, and who faces a different potential entrant in each market. It is assumed that the monopolist and the potential entrants are all risk neutral, and that the monopolist faces the  $T$  potential entrants in a sequential fashion. By the latter we mean that the monopolist first faces the potential entrant in what we refer to as market 1, and then sequentially faces the potential entrants in markets 2, 3...  $T$ . In each market, furthermore, payoffs to the players are as in Figure 1. Note, we consider the case  $0 < b < 1$  and  $a > 1$ . Also, it is assumed that when the potential entrant in each market  $k$  faces the monopolist, the potential entrant is aware of all moves by players in lower-numbered markets.

To complete the model, we must specify the manner in which players form expectations. With probability one, the monopolist

		PAYOFFS	
		ENTRANT'S	MONOPOLIST'S
ENTRANT STAYS OUT	MONOPOLIST ACQUIESCES	0	$a$
	MONOPOLIST FIGHTS	$b$	0
ENTRANT ENTERS		$b-1$	-1

FIGURE 1

is sophisticated and thus has rational expectations. On the other hand, for each market  $k$  there is a probability  $p$  that the potential entrant is sophisticated, and a probability  $(1-p)$  that the potential entrant is naive. If a potential entrant is sophisticated, then he also has rational expectations. If a potential entrant is naive, then he simply extrapolates from previous behavior. Specifically, if the last entry brought forth a cooperative (aggressive) response from the monopolist, then a naive potential entrant anticipates that with probability one a further entry will bring forth a cooperative (aggressive) response.<sup>13</sup> If there are no prior entries to extrapolate from, however, then a naive potential entrant anticipates that, with probability  $\pi$ , entry is met by cooperation and with probability  $(1-\pi)$  entry is met by aggression. Finally, to keep the analysis simple we place the following two restrictions on the parameters. First,  $\pi b + (1-\pi)(b-1) < 0$ , that is, a naive potential entrant does not enter if there are no prior entries to extrapolate from.<sup>14</sup> Second, for every integer  $k$  in the interval  $[1, T-1]$  it is the case that  $\sum_{i=1}^k (1-p)^i a \neq 1$ .

The following four propositions characterize perfect Nash equilibria for our model as a function of the exogenous parameters. By restricting the analysis to perfect Nash equilibria, we eliminate the possibility that sophisticated players maintain threats which they would not find rational to carry out (see Selten for a discussion of perfectness in this context).

<sup>13</sup>This type of oversensitivity by agents to the most recent information available is consistent with empirical evidence (see Arrow), and with the work of cognitive psychologists (see Kahneman et al.).

<sup>14</sup>This assumption is not critical, but rather as mentioned above serves to simplify the analysis. In our paper (1983), we indicate the results that follow from the alternative that  $\pi b + (1-\pi)(b-1) > 0$ .

<sup>12</sup>The description of Milgrom and Roberts above is somewhat imprecise. They actually show that the paradox can be resolved by having the potential entrants only perceive that there is a probability that the monopolist always acts aggressively.



**PROPOSITION 5:** *If  $p=1$  then (i) entry occurs in every market and (ii) the monopolist cooperates every time entry occurs.*

Proposition 5 tells us that in the absence of any limited rationality our model yields Selten's chain-store paradox. That is, when with probability one each potential entrant is sophisticated, entry occurs in each market and the monopolist cooperates each time entry occurs. We now analyze the other polar case.

**PROPOSITION 6:** *If  $p=0$ , then entry never occurs.*

Proposition 6 indicates that, when with probability one each potential entrant is naive, the model yields results exactly opposite from those that held when with probability one each potential entrant was sophisticated. Note, the proof of Proposition 6 depends on the assumption  $\pi b + (1-\pi)(b-1) < 0$ . However, the equilibrium would only change slightly with the opposite assumption. Specifically, if  $\pi b + (1-\pi)(b-1) > 0$ , equilibrium is characterized by the following three statements. First, the period 1 potential entrant would enter. Second, the monopolist would respond by acting aggressively. Third, entry would not occur in any other market.

We have now analyzed the two polar cases of pure rational expectations and pure limited rationality. Our next two propositions consider restrictions on the parameter space which are between these two polar cases.

**PROPOSITION 7:** *If  $\sum_{i=1}^{T-1}(1-p)^i a < 1$ , then (i) entry first occurs in the lowest-numbered market that contains a sophisticated potential entrant, (ii) entry occurs in every succeeding or higher-numbered market, and (iii) the monopolist cooperates each time entry occurs.*

Propositions 5 and 6 indicate that, for neither type of agent to be disproportionately important, entry should occur on average in  $pT$  markets. However, Proposition 7 indicates that when  $\sum_{i=1}^{T-1}(1-p)^i a < 1$ , entry occurs on average in  $\sum_{i=1}^{T-1}(1-(1-p)^i)$  markets, and it is easily shown that  $\sum_{i=1}^{T-1}(1-(1-p)^i)$  always exceeds  $pT$  and, in

fact, it frequently exceeds  $pT$  by a wide margin. For example, if  $T=10$  and  $p=1/2$ , then  $pT=5$  and  $\sum_{i=1}^9(1-(1-p)^i) \approx 9$ . Thus, for the parameterizations covered by Proposition 7, sophisticated agents are disproportionately important.

**PROPOSITION 8:** *If  $\sum_{i=1}^{T-1}(1-p)^i a > 1$ , then (i) entry does not occur in markets 1 through  $T-z$ , where  $z$  is the lowest integer for which  $\sum_{i=1}^z(1-p)^i a > 1$ , (ii) for markets numbered above  $T-z$ : (a) entry first occurs in the lowest-numbered market that contains a sophisticated potential entrant, (b) entry occurs in every succeeding or higher-numbered market, and (c) the monopolist cooperates each time entry occurs.*

As stated previously, Propositions 5 and 6 indicate that, for neither type of agent to be disproportionately important, entry should occur on average in  $pT$  markets. However, Proposition 8 tells us that, for some parameterizations, entry occurs on average in less than  $pT$  markets. For example, if  $(1-p)a > 1$ , then entry occurs on average in  $p$  markets, and  $p$  is obviously less than  $pT$ . Thus, Proposition 8 indicates that there are some parameterizations for which naive potential entrants have a disproportionately large effect on equilibrium.

Proposition 8 concludes our analysis of the chain-store paradox. In summary, we have found that, in a repeating-game situation where reputation is a potential factor, it is possible for either type of agent to dominate. That is, for some parameterizations sophisticated agents will have a disproportionately large effect on equilibrium, while for other parameterizations the naive agents will be the ones with the disproportionately large effect.

Finally, given that either type of agent can dominate, a question arises as to which type of agent will in general dominate in such repeated-game situations. Although it is difficult to quantify, we feel that, at least for chain-store-like games, the naive agents will in general be the dominant ones. There are two pieces of evidence that point in this direction. First, one interpretation of the Kreps-Wilson and Milgrom-Roberts results is that, when the probability for limited ra-

tionality is introduced on the side of the monopolist, then it is almost always the probability of naiveté that is disproportionately important. Second, in a richer version of the model analyzed in this section, the restriction needed for the naive agents to be dominant becomes much less stringent. Specifically, consider a model where there is some uncertainty concerning what the monopolist receives in a market when entry does not occur. Preliminary analysis suggests that as long as there is some probability that  $(1 - p)$  times this return is greater than one, the naive agents will be the ones who dominate.

#### IV. Conclusion

Papers that provide alternatives to the rationality assumption, as well as papers that contain the rationality assumption, have in general ignored the idea that agents tend to vary in terms of information processing abilities. Here we have attempted to investigate the ramifications of this type of heterogeneity. We did this by analyzing two simple models in which the population is composed of two groups. Agents in the first group were characterized by rational expectations and were referred to as sophisticated, while agents in the second group were characterized by incorrect expectations and were referred to as naive. The analysis yielded three major results. First, in a world characterized by congestion effects, sophisticated agents tend to have a disproportionately large effect on equilibrium. Second, in a world characterized by synergistic effects, naive agents tend to be disproportionately important. Third, in a repeating-game situation where reputation is a potential factor, it is possible for either type of agent to be dominant. Finally, the analysis also yielded insights concerning the interaction of informed and uninformed agents, where all agents have rational expectations. Specifically, the analysis suggests that, for situations which exhibit congestion effects, informed agents tend to have a disproportionately large effect on equilibrium. On the other hand, for situations that exhibit synergistic effects, the analysis suggests that uninformed agents are the ones who are disproportionately important.

The research presented in this paper could be extended in a number of directions. One direction of particular interest concerns the applicability of the results contained herein to the rational expectations challenge to Keynesian macroeconomics. This challenge has left macroeconomics in a state of flux. The reason being that, although the challenge has brought to the fore the theoretical inconsistencies inherent in Keynesian macroeconomics, simple rational expectations macroeconomic models do not yield predictions consistent with empirical observation. The profession has responded to this situation by embedding a variety of imperfections and rigidities into rational expectations models (for example, staggered long-term contracts are introduced in John Taylor, 1980, lags in the production process are introduced in Finn Kydland and Edward Prescott, 1982, and uncertainty with Bayesian learning is introduced in Roman Frydman and Edmund Phelps, 1983). The analysis in this paper suggests a different response. That is, rather than assuming rational expectations and then introducing imperfections, why not start with the more realistic assumption that agents are heterogeneous in terms of their information-processing abilities. We feel such an approach shows particular promise because of the following. First, given the similarity between our description of synergistic interaction among agents and old style Keynesian multiplier analysis, it would not be surprising if macroeconomic models with this type of heterogeneity exhibited synergistic effects. Second, our results suggest that if this is indeed the case, then it may very well be that the naive agents are the ones who would be dominant in such a model.<sup>15</sup>

#### APPENDIX

More detailed proofs of many of the propositions are contained in our earlier paper.

<sup>15</sup> See Akerlof and Yellen for macroeconomic examples wherein the presence of naive agents has a critical effect on the nature of equilibrium. Note, however, their results are not driven by the presence of either congestion or synergistic effects.

## PROOF of Proposition 1:

Suppose not, that is, suppose for instance that  $N_A^s > N_A^n$  and that there exists a  $0 < p < 1$ , denoted  $\hat{p}$ , such that  $N_A \leq \hat{p}N_A^s + (1 - \hat{p})N_A^n$ . Our suppositions that  $N_A^s > N_A^n$  and  $N_A \leq \hat{p}N_A^s + (1 - \hat{p})N_A^n$  together with the definitions of  $D$  and  $D^s$  yield that  $D$  at  $p = \hat{p}$  must exceed  $D^s$ . On the other hand, our two suppositions together with (3) and the definitions of  $D^n$ ,  $N_A^s$  and  $N_A^n$  yield that  $D^s$  must exceed or equal  $D$  at  $p = \hat{p}$ . Thus we have a contradiction, which means that if  $0 < p < 1$  and  $N_A^s > N_A^n$ , then  $N_A > pN_A^s + (1 - p)N_A^n$ . The other cases are demonstrated similarly.

## PROOF of Proposition 2:

Note, first, all the restrictions in the statement of the proposition are being assumed throughout the proof. We know that

$$(A1) \quad W = p \left[ \int_0^D (f_A(N_A) - c_i) g(c_i) dc_i + \int_D^C (f_B(N_B) - (C - c_i)) g(c_i) dc_i \right] + (1 - p) \left[ \int_0^{D^n} (f_A(N_A) - c_i) g(c_i) dc_i + \int_{D^n}^C (f_B(N_B) - (C - c_i)) g(c_i) dc_i \right].$$

Note,  $W = W^s$  when  $p = 1$ ,  $W = W^n$  when  $p = 0$ , and  $W = \hat{W}$  when  $D = D^s$ ,  $N_A = \hat{N}_A$  and  $N_B = \hat{N}_B$ , where  $\hat{N}_A = \int_0^{D^s} p g(c_i) dc_i + \int_0^{D^n} (1 - p) g(c_i) dc_i$  and  $\hat{N}_B = \int_{D^s}^C p g(c_i) dc_i + \int_{D^n}^C (1 - p) g(c_i) dc_i$ .

Consider a social welfare-maximizing government that could direct the actions of all the agents in the population. Such a government would choose  $D^*$  to maximize

$$(A2) \quad W^* = \int_0^{D^*} (f_A(N_A^*) - c_i) g(c_i) dc_i + \int_{D^*}^C (f_B(N_B^*) - (C - c_i)) g(c_i) dc_i,$$

where  $N_A^* = \int_0^{D^*} g(c_i) dc_i$

and  $N_B^* = \int_{D^*}^C g(c_i) dc_i$ .

Differentiation yields that  $W^*$  is a single-peaked function of  $D^*$ , that is,  $(\partial W^* / \partial D^*)$

$> 0$  for  $D^* < D^0$  and  $(\partial W^* / \partial D^*) < 0$  for  $D^* > D^0$ , where  $D^0 = \arg \max_{D^*} W^*$ .

Now consider a social welfare-maximizing government which could only direct the actions of the sophisticated agents in the population. Such a government would choose  $D^+$  to maximize

(A3)

$$W^+ = (1 - p) \left[ \int_0^{D^n} (f_A(N_A^+) - c_i) g(c_i) dc_i + \int_{D^n}^C (f_B(N_B^+) - (C - c_i)) g(c_i) dc_i \right] + p \left[ \int_0^{D^+} (f_A(N_A^+) - c_i) g(c_i) dc_i + \int_{D^+}^C (f_B(N_B^+) - (C - c_i)) g(c_i) dc_i \right],$$

where  $N_A^+ = \int_0^{D^n} (1 - p) g(c_i) dc_i + \int_0^{D^+} p g(c_i) dc_i$  and  $N_B^+ = \int_{D^n}^C (1 - p) g(c_i) dc_i + \int_{D^+}^C p g(c_i) dc_i$ . Differentiation yields that  $W^+$  is a single-peaked function of  $D^+$ . Denote  $N_A^o(N_B^o)$  as the number of agents who would wind up at path  $A(B)$ , given a government with this limited ability to direct behavior.

Suppose  $W^n < W^s$  and  $N_A^s > N_A^n$ . Given the single-peaked property of  $W^*$ , this implies  $N_A^n < N_A^s \leq N_A^o$ . Given Proposition 1, this in turn implies  $N_A > \hat{N}_A$ . It can be demonstrated that  $N_A^n, N_A^s < \hat{N}_A^o$  implies  $N_A < N_A^o$ .<sup>16</sup> Hence, we have  $N_A^o > N_A > \hat{N}_A$ . Given the single-peaked property of  $W^+$ , this in turn implies that if  $W^n < W^s$  and  $N_A^s > N_A^n$ , then  $W > \hat{W}$ . Moreover, given the symmetry of the problem, this result yields that  $W^n < W^s$  and  $N_A^s < N_A^n$  imply  $W > \hat{W}$ .

The assumptions  $f_A'' \leq 0$  and  $f_B'' \leq 0$  yield  $\hat{W} > \bar{W}$ . Given the preceding result, this in turn yields that  $W^s > W^n$  implies  $W > \bar{W}$ .

Finally, the proof that  $W^n > W^s$  implies  $W < \hat{W}$  follows along similar lines.

## PROOF of Proposition 3:

Suppose not, that is, suppose for instance that  $N_A^n < N_A^s$  and that there exists a  $0 < p <$

<sup>16</sup>See the mathematical supplement mentioned in fn. 7.

1, denoted  $\hat{p}$ , such that  $N_A \geq \hat{p}N_A^s + (1 - \hat{p})N_A^n$ . Given this, (3), and the definitions of  $N_A^n$  and  $N_A^s$  yield that  $D$  at  $p = \hat{p}$  must exceed or equal  $D^s$ . On the other hand, our assumption  $[f'_A(\int_0^z g(c_i)dc_i) + f'_B(\int_z^1 g(c_i)dc_i)]g(z)/2 < 1$  for all  $0 \leq z \leq C$  combined with (3) yields that  $N_A$  is a strictly increasing function of  $N_A^n$ . We also know  $N_A = N_A^s$  when  $N_A^n = N_A^s$ , which given the preceding implies  $N_A$  at  $p = \hat{p}$  must be strictly less than  $N_A^s$ . Furthermore, given  $f'_A \geq 0$ ,  $f'_B \geq 0$  and that one is always strictly positive, the fact that  $N_A^s$  exceeds  $N_A$  at  $p = \hat{p}$  yields a contradiction because it implies  $D^s$  exceeds  $D$  at  $p = \hat{p}$ . Thus, when  $N_A^n < N_A^s$ , it must be the case that  $N_A < pN_A^s + (1 - p)N_A^n$ . The other cases are demonstrated similarly.<sup>17</sup>

#### PROOF of Proposition 4:

Note, first, all the restrictions in the statement of the proof are assumed throughout the proof. Given these restrictions,  $W^*$  remains a single-peaked function of  $D^*$  and  $W^+$  a single-peaked function of  $D^+$ .

Suppose  $W^n < W^s$  and  $N_A^s > N_A^n$ . Given the single-peaked property of  $W^*$ , this implies  $N_A^n < N_A^s \leq N_A^0$ . In turn, given Proposition 3 and the single-peaked property of  $W^+$ , this yields that when  $W^n < W^s$  and  $N_A^s > N_A^n$ , then  $W < \hat{W}$ . Moreover, because of the symmetry of the problem, this result yields that  $W^n < W^s$  and  $N_A^n > N_A^s$  imply  $W < \hat{W}$ .

From the definitions, it is necessarily true that  $\hat{W} - \bar{W} = 0$  when  $N_A^s = N_A^n$ . It is also true that  $\partial(\hat{W} - \bar{W})/\partial N_A^n \geq 0$  when  $N_A^n \leq N_A^s$ . Together these two statements imply  $\hat{W} < \bar{W}$  when  $N_A^s \neq N_A^n$ . This, combined with

the preceding result, in turn, yields that  $W^s > W^n$  implies  $W < \bar{W}$ .

Finally, the proof that  $W^n > W^s$  implies  $W > \bar{W}$  follows along similar lines.

#### PROOF of Proposition 5:

Consider first market  $T$ . If entry occurs the only rational response for the monopolist is to cooperate. Given this, the only rational move for the market  $T$  potential entrant is to enter. Now consider market  $k$ , where it is known that entry will necessarily occur in every higher-numbered market. If entry occurs in market  $k$ , given that entry is going to occur in every higher-numbered market, the only rational response for the monopolist is to cooperate. This in turn yields that the only rational move for the market  $k$  potential entrant is to enter. Finally, the above two ideas yield that the only perfect Nash equilibrium is for entry to occur in every market, and for the monopolist to cooperate every time entry occurs.

#### PROOF of Proposition 6:

This follows immediately from our assumption  $\pi b + (1 - \pi)(b - 1) < 0$ .

#### PROOF of Proposition 7:

Consider market  $T$ . If entry occurs, the only rational response for the monopolist is to cooperate. Given this, the market  $T$  potential entrant will enter if he is sophisticated. Now consider market  $T - 1$ . Given the above and the fact that  $(1 - p)a < 1$ , the only rational response for the monopolist is again to cooperate. This in turn yields that the market  $T - 1$  potential entrant will enter if he is sophisticated. Furthermore, successively repeating this argument yields that entry will occur in each market that contains a sophisticated potential entrant, and the monopolist will cooperate every time entry occurs.

To complete the proof we need only demonstrate that a naive potential entrant will (will not) enter if there is (is not) a lower-numbered market that contains a sophisticated potential entrant. Consider first a naive potential entrant for whom there is a lower-numbered market that contains a sophisticated potential entrant. For this

<sup>17</sup>One point that should be noted is that this proof relies heavily on the assumption employed to eliminate the possibility of multiple equilibria. We conjecture that, without this assumption, there would always be equilibria consistent with the proposition. On the other hand, without the assumption, it would probably also be possible to have equilibria that were inconsistent with the proposition. What this suggests to us is that, when there are synergistic effects, it is possible for sophisticated agents to dominate. However, this would only occur when, in a sense, the presence of naive agents caused the economy to jump between different families of equilibria.

potential entrant, extrapolation yields that the monopolist cooperates, and thus such a potential entrant will enter. Now consider a naive potential entrant for whom there is no lower-numbered market that contains a sophisticated potential entrant. For this potential entrant, there will be no prior entries to extrapolate from, and thus such a potential entrant will not enter.

#### PROOF of Proposition 8:

Suppose (i) is true. Given this, (ii) follows from the same logic as in the proof of Proposition 7. Thus, we need only prove (i).

Consider a sophisticated potential entrant located in a market  $k$ , where  $1 \leq k \leq T - z$ . If this potential entrant were to enter, the monopolist would act aggressively because the expected return from the deterrence of future naive potential entrants would exceed the immediate return to cooperating, that is,  $\sum_{i=1}^{T-k} (1-p)^i a > 1$ . Thus, the only rational move for such a sophisticated potential entrant would be not to enter. Furthermore, given this, it follows immediately that any naive potential entrant located in a market  $k$ , where  $1 \leq k \leq T - z$ , would also decide not to enter.

#### REFERENCES

- Akerlof, George and Yellen, Janet, "The Macroeconomic Consequences of Near-Rational Rule-of-Thumb Behavior," *Quarterly Journal of Economics*, 1985 forthcoming.
- Arrow, Kenneth, "Risk Perception in Psychology and Economics," *Economic Inquiry*, January 1982, 20, 1-9.
- Conlisk, John, "Costly Optimizers Versus Cheap Imitators," *Journal of Economic Behavior and Organization*, 1980, 1, 275-93.
- Cooper, Russell and Ross, Thomas, "Prices, Product Qualities, and Asymmetric Information: The Competitive Case," *Review of Economic Studies*, April 1984, 60, 197-208.
- Cyert, Richard and DeGroot, Morris, "Rational Expectations and Bayesian Analysis," *Journal of Political Economy*, May/June 1974, 82, 521-36.
- and March, James, *A Behavioral Theory of the Firm*, Englewood Cliffs: Prentice-Hall, 1963.
- DeCanio, Stephen, "Rational Expectations and Learning from Experience," *Quarterly Journal of Economics*, February 1979, 93, 47-57.
- Frydman, Roman and Phelps, Edmund, *Individual Forecasting and Aggregate Outcomes: "Rational Expectations" Reexamined*, New York: Cambridge University Press, 1983.
- Grossman, Sanford and Stiglitz, Joseph, "The Impossibility of Informationally Efficient Markets," *American Economic Review*, June 1980, 70, 393-408.
- Hall, Robert L. and Hitch, Charles J., "Price Theory and Business Behavior," *Oxford Economic Papers*, May 1939, 2, 12-45.
- Haltiwanger, John and Waldman, Michael, "Rational Expectations and the Limits of Rationality: An Analysis of Heterogeneity," Working Paper No. 303, University of California-Los Angeles, December 1983.
- Kahneman, Daniel, Slovic, Paul and Tversky, Amos, *Judgement Under Uncertainty: Heuristics and Biases*, New York: Cambridge University Press, 1982.
- Keeler, Theodore and Small, Kenneth, "Optimal Peak-Load Pricing, Investment, and Service Levels on Urban Expressways," *Journal of Political Economy*, February 1977, 85, 1-26.
- Kreps, David and Wilson, Robert, "Reputation and Imperfect Information," *Journal of Economic Theory*, August 1982, 27, 253-79.
- Kydland, Finn and Prescott, Edward, "Time to Build and Aggregate Fluctuations," *Econometrica*, November 1982, 50, 1345-70.
- Machlup, Fritz, "Marginal Analysis and Empirical Research," *American Economic Review*, September 1946, 36, 519-54.
- Milgrom, Paul and Roberts, John, "Predation, Reputation, and Entry Deterrence," *Journal of Economic Theory*, August 1982, 27, 280-312.
- Nelson, Richard and Winter, Sidney, *An Evolutionary Theory of Economic Capabilities and Behavior*, Cambridge: Harvard University Press, 1982.
- Radner, Roy, "Satisficing," *Journal of Mathematical Economics*, June/September 1975, 2, 253-62.
- Russell, Thomas and Thaler, Richard, "The Rele-

- vance of Quasi Rationality in Competitive Markets," *American Economic Review*, September 1985, forthcoming.
- Selten, Richard, "The Chain-Store Paradox," *Theory and Decision*, April 1978, 9, 127-59.
- Simon, Herbert, "Theories of Decision-Making in Economics and Behavioral Science," *American Economic Review*, June 1959, 49, 253-83.
- Taylor, John, "Aggregate Dynamics and Staggered Contracts," *Journal of Political Economy*, February 1980, 88, 1-23.
- Williamson, Oliver, *Markets and Hierarchies: Analysis and Antitrust Implications*, New York: Free Press, 1975.

# Do Markets Differ Much?

By RICHARD SCHMALENSEE\*

This essay reports the results of a cross-section study of differences in accounting profitability that sheds light on some basic controversies in industrial economics. Most previous cross-section studies in this field have been concerned with testing hypotheses about structural coefficients in models meant to apply to essentially all markets. As we have learned more about the difficulties of constructing such general models and of performing tests on their structural parameters properly, structural cross-section analysis has fallen out of fashion. In contrast to most of the cross-section literature, the analysis reported here is fundamentally descriptive; it does not attempt directly to estimate or to test hypotheses about structural parameters.

I hope to show by example that one can perform illuminating analysis of cross-section data without a host of controversial maintained hypotheses. Cross-section data can yield interesting stylized facts to guide both general theorizing and empirical analysis of specific industries, even if they cannot

easily support full-blown structural estimation.<sup>1</sup> One can view the sort of search for stylized facts conducted here as either a replacement for or an input to interindustry structural estimation, depending on one's feeling about the long-run potential of that research approach. This study also departs from much of the cross-section literature by being fundamentally concerned with the *importance* of various effects, not just with coefficient signs and *t*-statistics.

In particular, this essay provides estimates of the relative importance of firm, market, and market share differences in the determination of business unit (divisional) profitability in U.S. manufacturing. Using 1975 data from the Line of Business Program of the U.S. Federal Trade Commission (FTC), I find support neither for the *existence* of firm effects nor for the *importance* of market share effects. Moreover, while industry effects apparently exist and are important, they appear to be *negatively* correlated with seller concentration in these data.

Section I relates firm, market, and share effects to current issues and controversies in industrial economics and thus supplies the motivation for our empirical analysis. The remainder of the essay treats the data and statistical methods employed (Section II), the empirical results obtained (Section III), and the main implications of those results (Section IV).

## I. Sources of Profitability Differences

In the *classical* tradition, following Joe Bain (1951, 1956), industrial economists treated the industry or market as *the* unit of

\*Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139. I am indebted to Stephen Postrel for excellent research assistance, to the FTC's Line of Business staff, particularly David Lean, William Long, and David Ravenscraft, for a variety of indispensable inputs, and to Ian Ayres, Richard Caves, Jerry Hausman, Paul Joskow, John Scott, anonymous referees, and, especially, Thomas Stoker for valuable advice. Seminar audiences at Stanford, Berkeley, Chicago, Northwestern, and British Columbia provided useful comments on earlier versions of this essay. Finally, I am grateful for financial support from the National Science Foundation, the U.S. Federal Trade Commission, and the Ford Motor Company (through a grant to MIT). The representations and conclusions presented herein are my own and have not been adopted in whole or in part by the FTC or its Bureau of Economics. The Manager of the Line of Business Program has certified that he has reviewed and approved the disclosure avoidance procedures used by the staff of the Line of Business Program to ensure that the data included in this paper do not identify individual company Line of Business data. I alone can be held responsible for this paper's contents.

<sup>1</sup>Christopher Sims (1980) has expressed a similar methodological position in the context of macroeconomics. The approach taken by Michael Gort and Rao Singamsetti (1976) is close in some respects to that taken here. They use firm-level data, however, and obtain very different results.

study. Differences among firms were assumed transitory or unimportant unless based on scale economies, which were generally found to be insubstantial. Equilibrium industry profitability was generally assumed to be primarily determined by the ability of established firms to restrict rivalry among themselves and the protection afforded them by barriers to entry. A central hypothesis in virtually all the classical work was that increases in seller concentration tend to raise industrywide profits by facilitating collusion. Most classical studies thus included concentration among the independent variables in regression analysis of industry average rates of return, and most published studies reported the coefficient of concentration to be positive and significant.<sup>2</sup>

An anticlassical, *revisionist* view of industrial economics has emerged in the last decade. In the simplest model consistent with this view, all markets are (at least approximately) competitive, and scale economies are absent (or negligible). The key assumption is that within at least some industries there are persistent efficiency differences among sellers.<sup>3</sup> Because more efficient enterprises tend both to grow at the expense of their rivals and to be more profitable, these differences tend to induce a positive intra-industry correlation between share and profitability even in the absence of scale economies. Moreover, the more important are efficiency differences in any industry, the less equal are market shares (and thus the higher is market concentration) and the higher are the profits of the leading firms (and thus the higher is industry average profitability). This model

thus predicts a positive correlation between concentration and profitability in cross section at the industry level even though, by assumption, concentration does not facilitate the exercise of market power.<sup>4</sup>

At the firm or (for multiproduct firms) business unit level, the revisionist view implies that market share should appear as the primary determinant of profitability in cross section regressions, while market concentration should have no impact. David Ravenscraft (1983) checked these predictions with FTC Line of Business data.<sup>5</sup> He found the impact of share on business unit profitability to be positive and highly significant, while the coefficient of concentration in the same regression was *negative* and significant. Ravenscraft interpreted his results as providing strong support for the revisionist argument that the significance of concentration in traditional industry-level cross-section regressions arises because concentration is correlated with share (and thus efficiency) differences, not because it facilitates collusion. Stephen Martin has recently obtained similar results in a simultaneous equations analysis of the FTC data. The strong relation between market share and profitability found by these and other authors is difficult to interpret within the classical tradition, given the apparent absence of important scale economies in most industries.<sup>6</sup>

A third tradition, which I will call *managerial*, has yet another set of implications for business unit profitability. Business

<sup>2</sup>Leonard Weiss (1974) provides a survey of cross-section studies in the classical tradition; see also F. M. Scherer (1980, ch. 9).

<sup>3</sup>Efficiency should not be interpreted in narrow process terms here. A product innovation may simply make a firm more efficient in the production of the Lancasterian characteristics it supplies to an existing market. While product innovations that yield true differentiation (by creating something approaching a new market) cannot be formally modeled in this fashion, it seems appropriate to think of nondramatic product innovations in efficiency terms for purposes of positive analysis of profitability.

<sup>4</sup>This new, revisionist view seems to have been articulated explicitly first by Harold Demetz (1973); see also Sam Peltzman (1977). Interesting formal models consistent with this view have recently been developed by Boyan Jovanovic (1982), S. A. Lippman and R. P. Rumelt (1982), and others. It is important to note that something like the classical notion of entry or mobility barriers (Richard Caves and Michael Porter, 1977) must be invoked to explain why imitation does not suffice to eliminate efficiency differences among firms in the revisionist model.

<sup>5</sup>Scherer (ch. 9) reviews earlier studies of the effects of market share. Most obtained results broadly consistent with those of Ravenscraft and Stephen Martin (1983) but used data sets apparently inferior to theirs.

<sup>6</sup>See Scherer (ch. 4) for an excellent survey of the available evidence on economies of scale.



schools and management consultants exist because it is widely believed that some firms are better managed than others and that one can learn important management skills that are not industry specific. In a widely acclaimed best seller, Thomas Peters and Robert Waterman, Jr. (1982) stress the importance of firm-level efficiency differences based in large measure on differences in "organizational cultures." Dennis Mueller (1977, 1983) has recently reported econometric results implying the existence of substantial, long-lived differences in measured firm profitability. When profit rates in 1950 are taken into account, Mueller (1983) finds that concentration has a significant *negative* coefficient in an equation explaining projected firm profit rates in 1972, and industry effects in general are relatively unimportant.

Both the revisionist and managerial alternatives to the classical tradition are based on plausible arguments and suggestive evidence. But I do not think that it has been shown that the classical attention to the industry was in any sense a mistake: case studies of real markets clearly reveal important differences. Why, then, do conventional market-level variables perform poorly or perversely when firm or share effects are included in cross-section regressions?

One probable reason comes readily to mind. It has long been recognized that we have very imperfect measures of the classic dimensions of market structure and basic conditions. Conditions of entry have proven particularly difficult to measure in a satisfactory fashion. Moreover, the link between the real, economic profitability dealt with in theoretical discussions and the accounting returns used in empirical work is weakened by inflation (Geoffrey Whittington, 1983), depreciation policy (Thomas Stauffer, 1971; Franklin Fisher and John McGowan, 1983), risk (myself, 1981), and both cyclical (Leonard Weiss) and secular (Ralph Bradburd and Richard Caves, 1982) disequilibria.<sup>7</sup>

Conventional, classical industry-level variables may thus perform poorly at least in part because they are poor, incomplete measures of the (classical and other) market effects present in available data. Since many of the usual classical industry-level variables are endogeneous in the long run, and it is difficult to formulate enough noncontroversial exclusion restrictions to identify all parameters of interest, it is not clear that problems of measurement and disequilibrium can be successfully attacked by structural modeling using available cross-section data.

## II. Methods and Data

Instead of attempting structural analysis, this study employs a simple analysis of variance framework that allows us to focus directly on the existence and importance of firm, market, and market share effects without having to deal simultaneously with specific hypotheses and measurement issues related to their determinants. Specifically, I deal in all that follows with the following basic *descriptive* model:

$$(1) \quad r_{ij} = \mu + \alpha_i + \beta_j + \gamma S_{ij} + \epsilon_{ij},$$

where  $r_{ij}$  is the (accounting) rate of return of firm  $i$ 's operations in industry  $j$ ,  $S_{ij}$  is its market share, the  $\alpha$ 's are firm effects, the  $\beta$ 's are industry effects,  $\mu$  and  $\gamma$  are constants, and the  $\epsilon$ 's are disturbances. The assumptions that market share enters linearly in (1) and that  $\gamma$  is the same for all industries are made mainly for comparability to the literature, though both also simplify computation and interpretation. The 1975 FTC Line of Business data set, which I use, contains information on large multidivisional firms. Such information is clearly required to separate firm and industry effects in (1).

While none of the coefficients in (1) can be given a defensible structural interpretation, analysis of that model as a whole can shed

<sup>7</sup>An additional accounting problem arises with business unit data: the allocation of shared assets among individual lines of business is inevitably somewhat arbitrary. If firms follow similar rules of thumb for doing

this, spurious industry effects can be added to business unit data.

light on the relative merits of at least the extreme versions of the classical, revisionist, and managerial positions. An extreme classical, for instance, would expect the  $\beta$ 's to differ substantially with  $\alpha_i = \gamma = 0$  for all  $i$ . Estimates consistent with these expectations would of course not exclude the possibility that industry effects simply reflect industry-wide differences between accounting and economic rates of return or industry-level disequilibria, with variations in monopoly power of little or no importance. But a finding that the  $\alpha$ 's and  $\gamma$  did not differ significantly from zero would cast doubt on extreme managerial or revisionist positions.

The implications of these last two positions for the parameters of equation (1) are slightly less clear cut. An extreme revisionist would presumably expect a large  $\gamma$  with all the  $\alpha$ 's and  $\beta$ 's near zero if the  $r_{ij}$  were observations on *equilibrium* rates of return. But, since our data are in fact for a single year and thus reflect the effects of cyclical and other short-run industry-level disequilibria, an extreme revisionist would not likely be surprised to find significant differences among the  $\beta$ 's estimated here. Similarly, an extreme managerial position might be that variations in the  $\alpha$ 's should be much more important in equilibrium than those in the  $\beta$ 's or in the  $\gamma S_{ij}$  terms. But an extreme managerialist would also not likely be surprised to find differences in the  $\beta$ 's in a single year's data. Moreover, firm-level efficiency differences might affect business unit profitability through the revisionist mechanism, so that firm-level and share effects might be hard to distinguish. (I investigate this possibility below.)

Using firm and industry dummy variables, I first use ordinary least squares (fixed effects estimation) and the usual  $F$ -statistics to test for the *existence* of market effects (nonidentical  $\beta$ 's), firm effects (nonidentical  $\alpha$ 's), and share effects (nonzero  $\gamma$ ) in (1) and the natural special cases thereof. To analyze the *importance* of these effects, I treat the actual  $\alpha$ 's,  $\beta$ 's,  $S$ 's, and  $\varepsilon$ 's in any particular sample as (unobservable) realizations of random variables with some joint population distribution. Under the usual assumption that  $\varepsilon$  is distributed independently of the other variables, the population variance of  $r$  can be

decomposed as follows:

$$\begin{aligned} (2) \quad \sigma^2(r) &= \sigma^2(\alpha) + \sigma^2(\beta) + \gamma^2 \sigma^2(S) \\ &\quad + \sigma^2(\varepsilon) + 2\rho(\alpha, \beta)\sigma(\alpha)\sigma(\beta) \\ &\quad + 2\gamma\rho(\alpha, S)\sigma(\alpha)\sigma(S) \\ &\quad + 2\gamma\rho(\beta, S)\sigma(\beta)\sigma(S), \end{aligned}$$

where the  $\rho$ 's are correlation coefficients and the  $\sigma$ 's are standard deviations. Depending on which effects are revealed to exist by the analysis of (1), I estimate either (2) or a special case thereof to provide information on the importance of the determinants of observed profitability. Estimates of (2) relate directly to the predictions of the alternative traditions discussed above. The particular (random effects) estimation techniques used in this phase of the analysis are presented in Section III.

In most of the statistical literature concerned with variance decomposition, orthogonality of effects is assumed, so that covariance terms like the last three on the right of (2) are set to zero.<sup>8</sup> But that assumption is not plausible here. If an important attribute of efficient firms is their ability to pick profitable industries in which to operate, for instance, we would expect this feature of the data generation process on which I must condition the estimates to produce a positive  $\rho(\alpha, \beta)$ . Similarly, one expects efficient firms to have low costs and high shares, so that  $\rho(\alpha, S)$  should be positive. Finally, if one knows that some particular  $S_{ij}$  is above average, one's conditional expectation must be that concentration in market  $j$  is above average. If one expects industry concentration to be positively related to industry profitability, it then follows that one expects  $\rho(\beta, S)$  to be positive. On the other hand, since  $\varepsilon$  captures all profitability differences unrelated to firm, industry, or market share differences, the assumption that it is orthogonal to those effects seems natural and reasonable.

The strength of this descriptive approach is that my conclusions about the three relevant types of effects will not be conditioned

<sup>8</sup>See, for instance, S. R. Searle (1971, chs. 9–11).

by maintained hypotheses regarding the determinants of those effects. I can focus directly on the general implications of extreme classical, revisionist, and managerial positions without having to deal with issues of endogeneity or identification. In addition, if one doubts a priori that any of these extreme positions is tenable, one can look to quantitative evidence on the importance of firm, market, and market share effects and the correlations among them to suggest tenable compromise positions as well as questions and strategies for future research.

One important issue of research strategy can be very easily addressed within this framework: is it defensible to work with industry-level data? Given the central role of profits in industrial economics, the answer must depend critically on how important industry effects are in determining industry rates of return. Only if industry profitability mainly reflects industry-level effects can one hope that hypotheses about the (classical, accounting, disequilibrium, and other) determinants of those effects can be productively tested with industry-level data. If  $R_j$  is the (appropriately weighted) average rate of return of business units operating in industry  $j$ , equation (1) implies

$$(3) \quad R_j = \mu + \beta_j + \text{terms in } \alpha\text{'s, } S\text{'s, and } \varepsilon\text{'s.}$$

Industry-level analysis would seem to be sensible if and only if (estimates of)  $\sigma^2(\beta)$  are large relative to the cross-section variance of the  $R_j$ , so that industry-level differences are important determinants of industry average rates of return.

All empirical results reported below are based on a subset of the 1975 data on individual business units gathered and compiled by the FTC's Line of Business Program. These business units account for about one-half of manufacturing sales and about two-thirds of manufacturing assets. (See Ravenscraft and the sources he cites for detailed discussions of the FTC data.) In order to minimize the influence of newly born and nearly dead operations, only the 3,816 business units present in the FTC data in both 1975 and 1976 were considered. Sixteen industries that appeared to be primarily residual classifications were excluded because they

seemed unlikely to correspond even approximately to meaningful markets.<sup>9</sup> This removed 340 observations. In order to mitigate scale-related heteroscedasticity problems and to focus on the revisionist mechanism (as distinguished from scale economies), the 1,070 remaining observations with market shares of less than 1.0 percent were excluded. (Note that none of these involve small firms; all are small divisions of the 471 large firms sampled by the FTC.) Finally, one outlier (with operating losses exceeding sales and assets/sales several times larger than other business units in its industry) was excluded before analysis began. Our final data set contained 1,775 observations on business units operated by 456 firms in 242 of the 261 FTC manufacturing industries.

In equation (1),  $r_{ij}$  was measured as the ratio of operating income to total assets, expressed as a percentage. This quantity provided an estimate of the total pre-tax rate of return (profits plus interest) on total capital employed; it seemed superior on theoretical grounds to the frequently employed price-cost margin as a measure of profitability.<sup>10</sup> Its mean was 13.66, and its variance,  $s^2(r)$ , was 348.97. For each industry in the sample, I also computed the asset-weighted average rate of return,  $R_j$ . The mean and variance of these 242 numbers were 13.08 and 86.91 ( $\equiv s^2(R)$ ), respectively. For  $S_{ij}$  I used estimates computed and kindly supplied by Ravenscraft.<sup>11</sup> The mean percentage market share in this sample was 6.14, with a variance of 59.23 ( $\equiv s^2(S)$ ).

### III. Empirical Findings

Figure 1 summarizes the results of least squares estimation of equation (1) and re-

<sup>9</sup>The industries dropped were the following: 20.29, 22.12, 23.06, 23.07, 24.05, 25.06, 28.17, 29.03, 30.06, 32.18, 33.13, 34.21, 35.37, 36.28, 37.14, and 39.08.

<sup>10</sup>Capital markets serve to equalize (risk-adjusted) rates of return on investment, not on sales. The case for using rate of return on sales as a measure of the Lerner index rests a belief that accounting average cost is a good proxy for marginal cost, which I doubt, and the undeniable proposition that sales are measured more accurately than assets.

<sup>11</sup>This variable is 100 times the variable *MS* used by Ravenscraft.

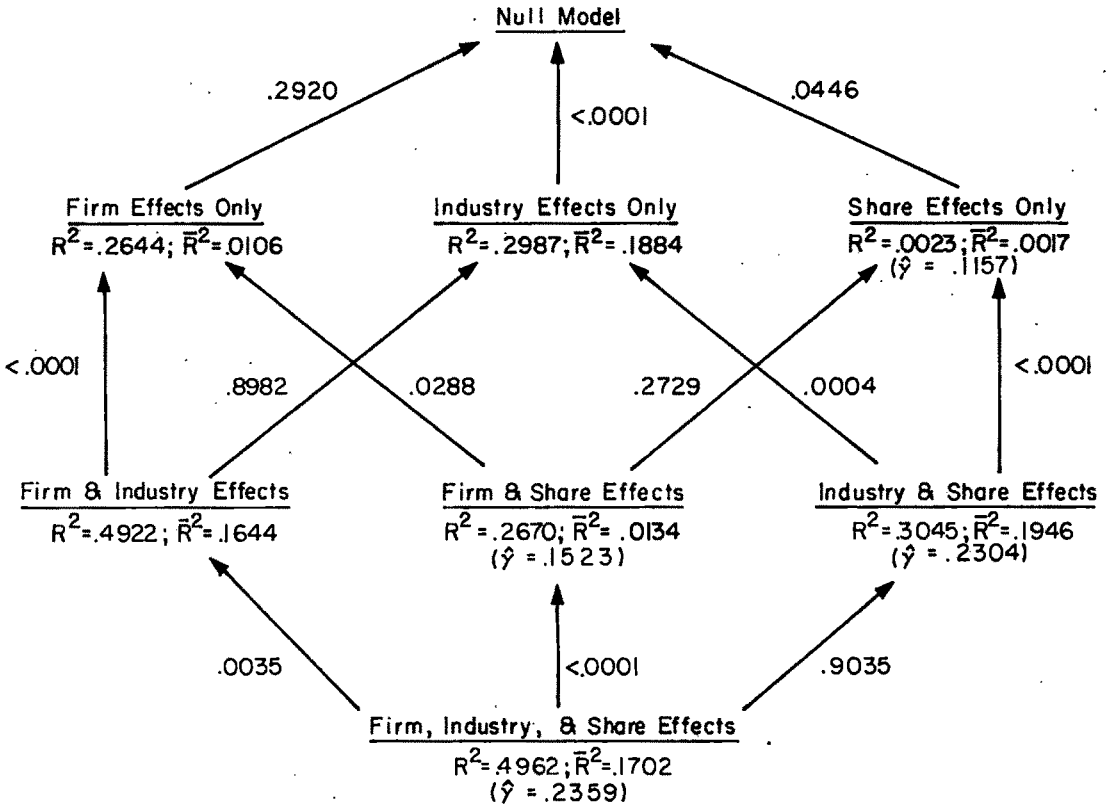


FIGURE 1

stricted models excluding one or more of the three effects with which we are concerned. The values of the ordinary and adjusted  $R^2$  statistics are shown,<sup>12</sup> along with the estimates of  $\gamma$  obtained from models with a share effect. Each arrow corresponds to the imposition of a restriction that one of the three effects discussed above is absent; the number next to each arrow is the probability level at which a standard  $F$ -test rejects that restriction. These numbers are referred to simply as  $P$ -levels in what follows.

All the high  $P$ -levels in Figure 1, which indicate failure to reject the null hypothesis at conventional levels, are generated by tests

for firm effects (arrows pointing to the right in Figure 1). These data imply that firm effects simply do not exist. In the absence of industry effects, the null hypothesis that the realized  $\alpha$ 's are identical can be rejected at the 29.2 percent level (no share effect) or the 27.3 percent level (share effect present). These results might lead a Bayesian analyst with a strongly managerial prior to accept the existence of firm effects. But both tests conducted in the presence of industry effects produce  $F$ -values less than unity, which provide absolutely no support for the existence of firm effects. Firm effects seem to approach significance only when firm-specific dummy variables serve as proxies for industry effects. When industry effects are controlled for, firm effects fade into insignificance. The absence of a similar interaction between firm and share effects indicates that firm effects do not operate through the revisionist mechanism to

<sup>12</sup>The adjusted  $R^2$  is equal to  $1 - [s^2(\epsilon)/s^2(r)]$ , where  $s^2$  is the usual unbiased estimator of the variance, so that changes in this quantity, rather than in  $R^2$  itself, correspond to changes in an unbiased estimator of the fraction of variance "explained."

any noticeable extent. Firm dummies do not serve as proxies for market share, and there is no difficulty disentangling firm and share effects.

In sharp contrast, all tests for the existence of industry or share effects produce significant results. All four tests of the null hypothesis of no share effects (arrows pointing to the left in Figure 1) signal rejection at  $P$ -levels below 4.5 percent, while the null hypothesis of no industry effects is always rejected at below the 0.01 percent level (vertical arrows in Figure 1).<sup>13</sup>

Let us now consider the importance of share and industry effects, postponing until Section IV a discussion of the implications of the absence of firm effects in these data. It is most instructive first to present an informal treatment based on information in Figure 1 and then to employ the relevant special case of (2) in a more systematic analysis.

Comparing adjusted  $R^2$ s of models not involving firm effects, market effects seem to account for between 18.84 and 19.29 percent of the sample variance of  $r$ . Following the discussion of equation (3), above, note that these percentages correspond to 75.65 and 77.46 percent of  $s^2(R)$ , the sample variance of industry average rates of return. Industry effects thus seem to be quite important, apparently accounting for the bulk of inter-industry differences in accounting rates of return. The industry seems an easily defensible unit of analysis.

On the other hand, the adjusted  $R^2$ s in Figure 1 indicate that market share effects add only between 0.17 and 0.62 percent to variance explained. Similarly, using  $\gamma = 0.2304$  from Figure 1,  $\gamma^2 s^2(S)$  amounts to only 0.90 percent of  $s^2(r)$ . It is interesting to note that in Ravenscraft's paper, which focuses on share effects, this ratio is even

smaller; it is between 0.53 percent ( $GLS$ ) and 0.82 percent ( $OLS$ ).<sup>14</sup> While Ravenscraft also uses 1975 Line of Business data, he uses the ratio of operating income to sales to measure profitability, does not delete "miscellaneous" industries or observations with small shares, uses classical variables like concentration in place of industry dummies, and attempts (in his  $GLS$  estimates) to correct for a complex pattern of heteroscedascity. The statistical significance but quantitative unimportance of market share effects thus seems a robust feature of these data.

One final pattern in the statistics presented above deserves mention. Market share adds more to adjusted  $R^2$  in the presence of industry effects (0.62 vs. 0.17 percent), and industry effects add more in the presence of share effects (19.29 vs. 18.84 percent). This sort of complementarity is suggestive of a negative correlation between market share and industry effects. Pointing in the same direction are the drops in the  $P$ -levels associated with share effects when industry effects are added and the corresponding changes (not visible in Figure 1) in the  $P$ -levels associated with industry effects. Finally, the fact that the estimate of  $\gamma^2 s^2(S)$  discussed above exceeds the contribution of share effects to adjusted  $R^2$  is also suggestive of a negative correlation between share and industry effects. (See equation (5) below.)

Let us now provide a more systematic analysis of the issues raised in the preceding three paragraphs. With no firm effects present, the relevant special cases of (1) and (2) are the following:

$$(4) \quad r_{ij} = \mu + \beta_j + \gamma S_{ij} + \varepsilon_{ij},$$

$$(5) \quad \sigma^2(r) = \sigma^2(\beta) + \gamma^2 \sigma^2(S) + \sigma^2(\varepsilon) \\ + 2\gamma\rho(\beta, S)\sigma(\beta)\sigma(S).$$

Readers uninterested in estimation technique and persuaded by the evidence presented above bearing on (5) may wish to glance briefly at Table 1, which summarizes the

<sup>13</sup> This is a very conservative statement of the strength of the evidence for the presence of industry effects. The  $F$ -statistics and corresponding restricted models are the following:  $F(241, 1533) = 2.709$ , null model;  $F(241, 1532) = 2.762$ , share effects only;  $F(241, 1078) = 2.007$ , firm effects only;  $F(241, 1077) = 2.033$ , firm and share effects. I calculate the probability of obtaining  $F$ s above any one of these values under the null hypothesis to be less than  $10^{-13}$ .

<sup>14</sup> The necessary statistics are in Tables 1 and A.1 of Ravenscraft.

TABLE 1—ESTIMATED VARIANCE DECOMPOSITIONS

Name	Population			Sample		
	Component	Estimate	Percentage	Component	Estimate	Percentage
Market	$\sigma^2(\beta)$	68.466	19.59	$\sigma^2(\beta)H$	67.905	19.46
Share	$\gamma^2\sigma^2(S)$	2.182	0.62	$\gamma^2\sigma^2(S)\{1-(1-H)\rho^2\}$	2.182	0.63
Covariance	$2\gamma\rho\sigma(\beta)\sigma(S)$	-2.177	-0.62	$2H\gamma\rho\sigma(\beta)\sigma(S)$	-2.159	-0.62
Error	$\sigma^2(\epsilon)$	281.049	80.41	$\sigma^2(\epsilon)$	281.049	80.54
Total	$\sigma^2(r)$	349.520	100.00	$s^2(r)$	348.977	100.00

Note: See text for sources and definitions. Totals may not add because of rounding.

results developed below, and then skip to Section IV.

Ordinary least squares estimation of (4), which appears in Figure 1 as the "Industry and Share Effects" model, yields a consistent and unbiased estimate of 281.05 for  $\sigma^2(\epsilon)$ . Following Searle's (chs. 9-11) treatment of variance components estimation in unbalanced models, I next compute consistent "analysis-of-variance" estimates of the remaining quantities on the right of (5).

Let the operator *ESS* mean "expected sum of squares about the sample mean," let  $N$  be the total number of observations, let  $N_j$  be the number of observations in industry  $j$ , and let  $M$  be the total number of industries. A bit of algebra yields

$$(6) \quad ESS(r_{ij} - \gamma S_{ij}) \\ = (N-1)\sigma^2(\epsilon) + (N-G)\sigma^2(\beta),$$

where

$$(7) \quad G = \sum_{j=1}^M (N_j)^2 / N.$$

If all industries had only one firm,  $G$  would equal one. If there were only one industry,  $G$  would equal  $N$ , since industry effects would not contribute to overall variance. In these data,  $G = 15.55$ . Using  $\gamma = .2304$  and  $\sigma^2(\epsilon) = 281.05$  from above, setting the expectation on the left of (6) equal to its sample value, and solving yields an estimate of 68.47 for  $\sigma^2(\beta)$ . This is equal to 19.62 percent of the sample variance of the  $r_{ij}$  and 78.78 percent of the sample variance of the  $R_j$ . The quantitative importance of industry effects

and the defensibility of industry-level analysis are again clear.<sup>15</sup>

In order to estimate the two remaining terms on the right of (5), it is necessary to be more specific about what is meant by a non-zero population correlation between market share and market effects. Imagine the data generation process first fixing the  $N_j$ , then drawing the  $\beta$ 's independently from their unconditional distribution, and finally drawing the  $S$ 's for each industry from the conditional distribution determined by the value of  $\beta$  previously drawn. Assume without loss of generality that the unconditional mean of the  $\beta$ 's is zero and of the  $S$ 's is  $\mu_s$ . I then impose the following assumptions:

(8a)

$$E(S_{ij}S_{kj}) = \begin{cases} (\mu_s)^2 + \sigma^2(S), & i = k \\ (\mu_s)^2 + \sigma^2(S)\rho^2(\beta, S) & i \neq k \end{cases}$$

$$(8b) \quad E(\beta_j S_{ij}) = \rho(\beta, S)\sigma(\beta)\sigma(S).$$

The first part of (8a) and (8b) are not restrictive; the second part of (8a) is consistent with but does not impose normality. These expectations are taken with respect to the unconditional population distribution,

<sup>15</sup>As a final check on the robustness of this conclusion, I computed MIVQUEO estimates of orthogonal firm, market, and error variance components of  $r_{ij}$  and  $(r_{ij} - \gamma S_{ij})$ . (See H. O. Hartley et al., 1978.) I obtained estimates of  $\sigma^2(\beta)$  of 62.03 and 64.88, respectively. This very different technique thus produced estimates very close to those in the text, further strengthening the case for the quantitative importance of industry effects in these data.

but they are *conditional* on the assignment of firms to markets. Similarly, for  $h \neq j$ ,  $E(\beta_h \beta_j) = E(\beta_h S_{ij}) = 0$ , and  $E(S_{ih} S_{kj}) = (\mu_s)^2$ .

Let  $r_j$  be the *unweighted* mean of the rates of return of business units in industry  $j$ . Then if (4) is the true model, (8a) and some algebra yield

$$(9) \quad ESS(r_{ij} - r_j) = (N - M)\gamma\sigma^2(S) \\ \times [1 - \rho^2(\beta, S)] + (N - M)\sigma^2(\varepsilon).$$

The quantity on the left is the expected sum of squared residuals from a regression of the  $r_{ij}$  on  $M$  industry dummy variables. This regression appears as the "Industry Effects Only" model in Figure 1. Use of (8) and a bit more algebra yields

$$(10) \quad ESS(r_{ij})/(N - 1) = E[\sigma^2(r)] \\ = H\sigma^2(\beta) + \gamma^2\sigma^2(S)[1 - (1 - H)\rho^2(\beta, S)] \\ + \sigma^2(\varepsilon) + 2H\gamma\rho(\beta, S)\sigma(\beta)\sigma(S),$$

where

$$(11) \quad H = (N - G)/(N - 1).$$

Equation (10) provides a decomposition of the *sample* variance of business unit profitability corresponding to the decomposition of the *population* variance given by (5).

Setting expectations equal to sample values, solving (9) for  $\gamma\sigma(S)$  and substituting into (10), an equation is obtained involving  $\rho(\beta, S)$ , sample statistics, and estimates derived above. A search of the interval  $(-1, +1)$  reveals a unique root; the estimated value of  $\rho(\beta, S)$  is  $-0.089$ . This confirms the negative correlation between industry and share effects. Equation (9) then yields an estimate of 2.182 for  $\gamma^2\sigma^2(S)$ . As this is only about 3.2 percent of the estimated value of  $\sigma^2(\beta)$ , the unimportance of share effects is also confirmed. Table 1 reports the estimated population and sample decompositions, corresponding to equations (5) and (10), respectively, implied by these estimates.

#### IV. Conclusions and Implications

The analysis of Section III indicates that the 1975 FTC Line of Business data provide strong support for the following four empirical propositions:

PROPOSITION 1: *Firm effects do not exist.*

PROPOSITION 2: *Industry effects exist and are important, accounting for at least 75 percent of the variance of industry rates of return on assets.*

PROPOSITION 3: *Market share effects exist but account for a negligible fraction of the variance of business unit rates of return.*

PROPOSITION 4: *Industry and market share effects are negatively correlated.*

The apparent nonexistence of firm effects is somewhat surprising. This finding is perfectly consistent with substantial intra-industry profitability differences, which Table 1 shows to be present in these data. The absence of firm effects in (1) merely means that knowing a firm's profitability in market  $A$  tells nothing about its likely profitability in randomly selected market  $B$ . This is consistent with the conglomerate bust of the past decade and with a central prescriptive thrust of Peters and Waterman (ch. 10): wise firms do not diversify beyond their demonstrated spheres of competence. The nonexistence of firm effects suggests that Mueller's (1983) persistent firm-level profitability differences are traceable to persistent differences at the business unit or industry level, combined with relatively stable patterns of activity at the firm level.<sup>16</sup>

The finding that industry effects are important supports the classical focus on industry-level analysis as against the revisionist tendency to downplay industry differences.

<sup>16</sup> Using firm and industry dummy variables to analyze FTC Line of Business data, John Scott (1984) finds significant firm effects on R&D intensity. This finding indicates that the absence of firm effects is not in any sense built into the FTC data. Since R&D spending reflects policy rather than performance, it is not surprising that firm effects show up there but not here.

But it is important to note that my analysis is generally silent on the merits of classical models and hypotheses. The empirical analysis here is basically descriptive, not structural. The results cannot exclude the possibility that industry-level profitability differences in 1975 were dominated by the effects of the severe recession and energy price shocks that were buffeting the economy. Though 1975 was an outlier in several respects, these considerations should serve to remind us that cyclical and other short-run disequilibrium effects are present in observed rates of return in any single year cross section. And there is no reason to suppose that such effects are uncorrelated with classical industry-level variables. This study cannot be interpreted as supporting an uncritical return to classical cross-section regressions. Finally, it is important to recognize that 80 percent of the variance in business unit profitability is unrelated to industry or share effects. While industry differences matter, they are clearly not all that matters.

The statistical significance of market share in my fixed-effects regressions is consistent with previous studies that have reached revisionist conclusions. I depart from those studies by directly examining the importance of market share in explaining variations in business unit profitability. My finding that share matters but doesn't matter much might seem to justify ignoring the revisionist mechanism in future research and policymaking. I think that would be a mistake.

First, the estimated coefficient of market share is quite large in equations with industry dummy variables. The "Industry and Share Effects" estimate of  $\gamma = .2304$  reported in Figure 1 implies that an increase of market share from 10 to 50 percent is on average associated with an increase of 9.2 percentage points in  $r_{ij}$ . Average profitability differences of this magnitude cannot sensibly be ignored, whatever their cause.

Second, even if the revisionist mechanism is unimportant on average in explaining profitability differences, it may be of central importance in some markets or classes of markets. The coefficient of share is constrained to be equal across industries in our regressions, even though a large number of authors have found substantial and (to some

extent) systematic differences in the profitability/share relation across industries. Mueller (1983), for instance, finds the coefficient of market share in a profitability equation rises in cross section with increases in industry advertising intensity, which he interprets as reflecting basic conditions that make possible product differentiation.<sup>17</sup> The basic revisionist mechanism seems too plausible to dismiss entirely; we ought instead to investigate the *industry-level* factors that affect its nature and importance.

Finally, the negative correlation between market share and industry effects is surprising indeed. Since concentration and market share are positively correlated, this finding is perfectly consistent with the negative concentration coefficients obtained by Ravenscraft, Martin, and Mueller (1983) in cross-section profitability regressions. Moreover, my results imply that those coefficients cannot be made to change sign by obvious respecification along classical lines.

One plausible explanation for negative concentration coefficients that also applies to negative values of  $\rho(\beta, S)$  has been advanced by Martin. He argues that capital-intensive, concentrated industries were hit hardest by recession and energy shocks in 1975 and that these same disequilibrium effects swamped any long-run effects of concentration on collusion. Note, however, that Ravenscraft finds that concentration has a *positive* sign in industry-level profitability regressions with these same data. At the very least, all this suggests the value of gathering and using panel data that would permit explicit analysis of cyclical and secular disequilibria.

<sup>17</sup>See also Caves and Thomas Pugel (1980), William Comanor and Thomas Wilson (1974, ch. 10), Allan Daskin (1983), and Porter (1979).

## REFERENCES

- Bain, Joe S., "Relation of Profit Rate to Industry Concentration: American Manufacturing, 1936-1940," *Quarterly Journal of Economics*, August 1951, 65, 293-324.
- , *Barriers to New Competition*, Cambridge: Harvard University Press, 1956.



- Bradburd, Ralph M. and Caves, Richard E., "A Closer Look at the Effect of Market Growth on Industries' Profits," *Review of Economics and Statistics*, November 1982, 64, 635-45.
- Brozen, Yale, *Concentration, Mergers, and Public Policy*, New York: Macmillan, 1982.
- Caves, Richard E. and Porter, Michael E., "From Entry Barriers to Mobility Barriers: Conjectural Decisions and Contrived Deterrence to New Competition," *Quarterly Journal of Economics*, May 1977, 91, 241-61.
- \_\_\_\_\_, and Pugel, Thomas E., *Intraindustry Differences in Conduct and Performance: Viable Strategies in U.S. Manufacturing Industries*, New York: New York University Graduate School of Business Administration, Salomon Brothers Center for the Study of Financial Institutions, 1980.
- Comanor, William S. and Wilson Thomas A., *Advertising and Market Power*, Cambridge: Harvard University Press, 1974.
- Daskin, Allan J., "Essays on Firm Diversification and Market Concentration," unpublished doctoral dissertation, MIT, November 1983.
- Demsetz, Harold, "Industry Structure, Market Rivalry, and Public Policy" *Journal of Law and Economics*, April 1973, 16, 1-10.
- Fisher, Franklin M. and McGowan, John J., "On the Misuse of Accounting Rates of Return to Infer Monopoly Profits," *American Economic Review*, March 1983, 73, 82-97.
- Gort, Michael and Singamsetti, Rao, "Concentration and Profit Rates: New Evidence on an Old Issue," *Explorations in Economic Research*, Winter 1976, 3, 1-20.
- Hartley, H. O., Rao, J. N. K. and Lamotte, Lynn, "A Simple Synthesis-Based Method of Variance Components Estimation," *Biometrics*, June 1978, 34, 233-42.
- Jovanovic, Boyan, "Selection and the Evolution of Industry," *Econometrica*, May 1982, 50, 649-70.
- Lippman, S. A. and Rumelt, R. P., "Uncertain Imitability: An Analysis of Interfirm Differences in Efficiency under Competition," *Bell Journal of Economics*, Autumn 1982, 13, 418-38.
- Martin, Stephen, *Market, Firm, and Economic Performance*, New York: New York University Graduate School of Business Administration, Salomon Brothers Center for the Study of Financial Institutions, 1983.
- Mueller, Dennis C., "The Persistence of Profits above the Norm," *Economica*, November 1977, 44, 369-80.
- \_\_\_\_\_, *The Determinants of Persistent Profits*, Washington: Bureau of Economics, U.S. Federal Trade Commission, June 1983.
- Peltzman, Sam, "The Gains and Losses from Industrial Concentration," *Journal of Law and Economics*, October 1977, 20, 229-63.
- Peters, Thomas J. and Waterman, Robert H., Jr., *In Search of Excellence: Lessons from America's Best-Run Companies*, New York: Harper & Row, 1982.
- Porter, Michael E., "The Structure within Industries and Companies' Performance," *Review of Economics and Statistics*, May 1979, 61, 214-28.
- Ravenscraft, David J., "Structure-Profit Relationships at the Line of Business and Industry Level," *Review of Economics and Statistics*, February 1983, 65, 22-31.
- Scherer, F. M., *Industrial Market Structure and Economic Performance*, 2d ed., Chicago: Rand-McNally, 1980.
- Schmalensee, Richard, "Risk and Return on Long Lived Tangible Assets," *Journal of Financial Economics*, June 1981, 9, 185-205.
- Scott, John T., "Firm versus Industry Variability in R&D Intensity," in Zvi Griliches, ed., *R&D, Patents, and Productivity*, Chicago: University of Chicago Press, 1984, 233-45.
- Searle, S. R., *Linear Models*, New York: Wiley & Sons, 1971.
- Sims, Christopher, "Macroeconomics and Reality," *Econometrica*, January 1980, 48, 1-48.
- Stauffer, Thomas R., "The Measurement of Corporate Rates of Return: A Generalized Formulation," *Bell Journal of Economics*, 2, Autumn 1971, 434-69.
- Weiss, Leonard W., "The Concentration-Profit Relationship and Antitrust," in Harvey J. Goldschmid et al., eds., *Industrial Concentration: The New Learning*, Boston: Little-Brown, 1974.
- Whittington, Geoffrey, *Inflation Accounting: An Introduction to the Debate*, Cambridge: Cambridge University Press, 1983.

# A Theory of Contractual Structure in Agriculture

By MUKESH ESWARAN AND ASHOK KOTWAL\*

Agricultural tenancy will continue to fascinate economists until the main riddle it poses is satisfactorily resolved. The challenge is to explain not only the existence of sharecropping, but also the existence of fixed rental and fixed wage contracts. The dominant contractual form can vary with the crop, the prevailing technology, the extent of market development, and other characteristics of the economic and social environment. Sometimes two or all three types of contracts seem to coexist in the same area. A general framework that could give convincing explanations of not only the existence of all three types of contracts but also of their spatial and temporal variation, and thus of their links with the economic environment, has yet to emerge.

There are three types of explanations that have been offered for the existence of different tenurial contracts:<sup>1</sup> (i) tradeoff between risk sharing and transaction costs, (ii) screening of workers of different qualities, and (iii) market imperfections for inputs besides land.

Steven Cheung (1969) postulated that sharecropping offers the advantage of risk sharing while the other two contracts involve lower transaction costs. The optimal tradeoff in a given set of circumstances would then determine the dominant contractual form.<sup>2</sup>

There are two difficulties with this view. First, the argument that risk sharing is the main motivation behind sharecropping lacks empirical support (Chandrashekar Pant, 1981; C. H. Hanumantha Rao, 1971; K. Chao, 1983). Second, the tradeoff argument lacks credibility as there is no reason to believe that sharecropping involves greater transaction costs than a wage contract if transaction costs include supervision costs, as they should.<sup>3</sup>

William Hallagan (1978) and David Newbery-Joseph Stiglitz have used screening models to explain the existence of different contracts.<sup>4</sup> The main problem warranting sharecropping is envisaged to be the asymmetry of information about tenants' abilities; different contracts are needed to sort tenants out by their abilities. We believe that the assumption of ignorance on the part of landlords about tenants' abilities is quite inappropriate for most rural communities. Typically, there is little mobility, thus information about abilities and assets is easily available. Moreover, screening models cannot explain why a certain contractual form might predominate in one area while quite another form predominates elsewhere.<sup>5</sup> Just as the coexistence of contracts (if present) needs to be explained, a lack of coexistence also requires explanation. Screening models

\*Department of Economics, University of British Columbia, Vancouver, B.C., V6T 1Y2 Canada. We are grateful to Robert Allen, Chris Archibald, Charles Blackorby, Jean Dreze, Curtis Eaton, John Harris, Tracy Lewis, Michael Manove, Michel Patry, Craig Riddell, Pankaj Tandon, John Weymark, Ralph Winter, the participants of Economic Theory workshops at the University of British Columbia and Simon Fraser University, and anonymous referees for helpful comments.

<sup>1</sup>For a comprehensive recent survey of tenancy models, see Hans Binswanger and Mark Rosenzweig (1982). The present discussion of the existing literature borrows heavily from it.

<sup>2</sup>Cheung and Joseph Stiglitz (1974) had considered production uncertainty as the only source of risk. But in response to the criticism that mixing wage and fixed

rental contracts would offer the same risk-sharing advantage as sharecropping, the more recent literature has considered other sources of risk (for example, David Newbery and Stiglitz, 1979), restoring the risk-sharing advantage to sharecropping. Newbery and Stiglitz point out that economies of scale would serve the same function.

<sup>3</sup>See Gerald Jaynes (1982) for a detailed critique of Cheung's argument. Also see Joseph Reid (1976).

<sup>4</sup>Franklin Allen (1982) uses a similar screening model that incorporates the unobservability of the quality of land in addition to the unobservability of the tenant's ability.

<sup>5</sup>For example, the dominant tenancy form in Japan was fixed rental while in East India it is sharecropping.

are also not capable of explaining the change in the contractual structure that has been observed to result from a change in technology or the development of markets. For example, tenancy contracts changed to wage contracts in India after the introduction of new technology in the 1960's (Rao, 1977, ch. 12); the sharecropping contracts in the post-bellum South of the United States changed to wage contracts with the advent of mechanization (Richard Day, 1967).

The most appealing view of tenancy is that it substitutes for the absence or imperfections of a market for some factor input besides land. The absence or incompleteness of markets can typically result from the high costs of quality enforcement. Recent literature has pointed out technical know-how (Joseph Reid, 1976), managerial ability (Clive Bell-Pinhas Zusman, 1979), bullocks (Christopher Bliss-Nicholas Stern, 1982), credit (Gerald Jaynes, 1982), and family labor (Pant, 1983) as examples of factors for which markets are highly imperfect. An effective way of gaining access to such a factor is to offer a self-monitoring (incentive) contract to the factor owner, involving him in the production process. The factor input is thus available only as a package deal with the factor owner's time. However, the self-monitoring contract does not have to be a share contract. The landlord could gain access to the tenant's supervision ability or to his bullocks by offering him a fixed rental contract. Why then does sharecropping exist?

Following Reid (1977) we envisage the landlord and tenant as both contributing unmarketed resources in a sharecropping arrangement. We view sharecropping as a partnership arrangement in which both agents have incentives to self-monitor.<sup>6</sup> Such a contract arises to mitigate morally hazardous behavior on the part of both agents—a phenomenon as yet unexplored in the literature. If all the monitoring of input qual-

ity is undertaken by a single agent, he becomes the sole residual claimant; in a wage contract it is the landlord, and in a fixed rental contract it is the tenant. The different contracts thus reflect different techniques of combining unmarketed productive inputs.<sup>7</sup> The choice of technique depends on exogenous parameters such as the endowment distribution across the classes of factor owners and the prevailing production technology. The equilibrium contractual structure emerges from the optimizing decisions of both landlord and tenant in a given environment. In order to facilitate the derivation of the testable implications of this view, we shall abstract from all considerations pertaining to risk in this paper. This abstraction enables us to make a comparison of the implications of the hypothesis that sharecropping exists to pool unmarketed resources with the implications of the conventional wisdom that it exists to share risk.

The paper is organized as follows. In Section I we present the general model of contractual choice in an agrarian economy. In Section II we solve the model and present comparative static results for a Cobb-Douglas specification. We use these results to explain some real world observations pertaining to agrarian contractual structures and changes therein. In Section III we discuss the limitations of the model and suggest extensions.

### I. The Model

Here we spell out our model of agricultural production that will enable us to endogenize the type of contractual arrangement (fixed wage, fixed rental, or sharecropping) that will prevail in a given setting. The model we construct is the simplest one that incorporates what we envisage to be the crucial features of agricultural production. It is our view that agricultural production entails the

<sup>6</sup> Bliss and Stern (p. 309) also view sharecropping as an arrangement that involves the pooling of managerial and cultivating skills. Peter Murrell (1983) points out that in the Philippines the word for sharecropping also means partnership.

<sup>7</sup> This idea is implicit in Reid's work on sharecropping (1977; 1976). In Reid (1979) this view is made explicit. We formalize the idea and relate it to the exogenous parameters of the social and economic environment.

use of certain unmarketed resources, and access to these resources is obtained by making their owners residual claimants. We believe this to be the most important determinant of the contractual structure of an agrarian economy and have therefore made it the cornerstone of our model of agrarian contracts. We focus on two specific unmarketed resources: the ability to supervise labor; and the managerial ability to make production decisions based on technical know-how and market information. These two abilities constitute the core of farm management.

Labor supervision is of crucial importance because in farming the output is very sensitive to the quality of effort; a slight mistiming in transplanting or the application of a wrong fertilizer mix, for example, can have disastrous consequences. Also, many farming activities are characterized by the fact that the quality of effort applied cannot be easily ascertained until after the work has been completed. The importance of the quality of effort on the one hand, and the inherent moral hazard problem of shirking on the other, imply that labor of itself is not an effective input; it is an aggregate of the labor hired and the supervision effort applied to reduce shirking that is an effective input into agricultural production. If  $L$  denotes the amount of labor hired and  $s$  the amount of time spent by a supervisor on supervising labor, we may define the "effective labor" input or "effort,"  $E$ , by

$$(1) \quad E = g(s, L; \epsilon),$$

where  $g$  is a linearly homogeneous aggregator which is increasing and concave in  $s$  and  $L$ , and  $\epsilon$  is a parameter ( $0 \leq \epsilon \leq 1$ ) characterizing the aggregator. The parameter  $\epsilon$  is introduced to capture the relative importance of supervision in a unit of effective labor, and is defined to be such that if  $\epsilon = 0$ , supervision is redundant, while increasing  $\epsilon$  implies that supervision is increasing in importance. An improvement in the technology of labor supervision would render supervisory effort less important and would therefore translate into a decrease in  $\epsilon$ .

In a tenancy contract, family labor is regarded as a crucial resource (Pant, 1983). Since it is easier to supervise one's own family labor than to supervise hired workers, a tenant may be considered to have labor-supervision abilities superior to those of a landlord. A tenancy contract should thus be considered not just as a contract to hire the labor of the tenant but, as well, the supervised labor of the family.

The efficiency of agricultural production is also crucially dependent on the quality of management decisions. Judicious choice of crops, proper land and water management, selection of inputs, and timely procurement and application of inputs are essential for successful farming. It involves decision making based on sound technical and market information. The choice of crops, for example, should depend on the expected changes in relative prices; water distribution should depend on the requirements of the specific technology; the choice and procurement of inputs should require the knowledge of the available inputs, their quality and prices, and the expected supply bottlenecks and shortages. Good production decisions must also depend on the knowledge of government tax-subsidy programs and of agricultural policy in general. Collecting such information and checking its reliability is a time-consuming process. The greater the time spent in gathering and assimilating information, the better is the quality of management decisions.<sup>8</sup> We shall, therefore, use the time devoted to such activities by the farm manager as a proxy for the managerial input.

We thus posit that agricultural production entails the use of four inputs: (i) management, which we represent by the time,  $t$ , spent by the entrepreneur on the activity; (ii) materials,  $M$ , which is an aggregate of such physical inputs as seeds, fertilizer fixed capital, etc.; (iii) labor effort,  $E$ , as defined in (1) above; and (iv) land,  $H$ , which we

<sup>8</sup> Bliss and Stern give an excellent account of the different managerial tasks important in agricultural production.

assume to be fixed and indivisible.<sup>9</sup> The output,  $q$ , of a farm may thus be written

$$(2) \quad q = \theta F(t, M, E, H),$$

where  $F$  is assumed to be a production function that is linearly homogeneous, increasing, and concave in its arguments. In (2),  $\theta$  is a positive random variable with an expected value of unity, intended to embody the effects of such stochastic factors as weather. Since we are explicitly abstracting from considerations pertaining to risk, the only role played by  $\theta$  is that it renders impossible the imputation of the amount of an input applied from knowledge of the levels of the other inputs and of the output. While the amounts of hired labor and material inputs are easily observable, supervision and management are not. This feature introduces the moral hazard problem of shirking in the provision of the supervision and management inputs in the event that they are provided by separate agents—as in sharecropping.

Substituting for  $E$  from (1), (2) may be rewritten as

$$(3) \quad q = \theta f(t, s, M, L, H; \epsilon),$$

where  $f$  is linearly homogeneous, increasing, and concave in its first five arguments.

In traditional societies, many exchange relationships are personalized and access to market information, scarce inputs, and government services is a matter of privilege. This privilege may stem from wealth, social position, or family connections. Rich landlords often dominate the local political bodies.<sup>10</sup> Access of any kind is correlated with the size of landholdings. Capital markets are usually

underdeveloped and access to credit is notoriously dependent on the ability to offer collateral, mostly land (Rao, 1977, pp. 138–42). In view of the limited ability of the marginal farmers and tenants to raise working capital, it is natural that it is the landlords who normally interact with traders and financial institutions. Landlords thus develop better avenues for the acquisition of productive market information—the government bureaucracy and the traders. In addition to these factors, landlords' families, with their higher wealth and social standing, are likely to have acquired better education than the laborers' families (Rao, 1977, p. 138). Landlords as a class, therefore, are more proficient in acquiring information on changing market conditions and technology than their tenants (Shigeru Ishikawa, 1981, pp. 168; Reid, 1979, p. 307; Lee Alston-Robert Higgs, 1982, p. 335). Resident landlords who are not alienated from farming and the rural scene are, therefore, likely to be much better suited for the role of overall decision making (i.e., general manager), while the tenant with his family labor force is better suited for the role of a labor supervisor (i.e., foreman).

For convenience we shall refer to the input  $t$  in (3) as "management" and to  $s$  as "supervision." We have argued above that landlords have superior abilities in management and tenants in supervision. We quantify this idea of differential abilities by means of two parameters,  $\gamma_1$  and  $\gamma_2$ . We assume that one hour of the landlord's (tenant's) time devoted to supervision (management) is equivalent to only a fraction  $\gamma_1$  ( $\gamma_2$ ) of one hour devoted to supervision (management) by the tenant (landlord). (Note that here and elsewhere in the paper, quantities subscripted by 1 refer to the landlord, those by 2 refer to the tenant.) An efficiency unit of management (supervision) will be assumed to be one hour of the landlord's (tenant's) time. Since market development tends to equalize access to know-how across agents by diffusing information, it would increase  $\gamma_2$  towards 1. Similarly, a substitution of the landlord's family labor for hired labor would increase  $\gamma_1$  towards 1 by eliminating the supervision advantage of tenants.

<sup>9</sup>It would be more realistic to make the plot size a choice variable controlled by the landlord and have a landlord enter into contracts of one of more variety with several tenants simultaneously. However, such a modification would greatly complicate the task of theorizing, without necessarily adding any new insights into the determination of contractual structure.

<sup>10</sup>See Rao (1977, pp. 186–87) for a discussion of the social and political power wielded by the rural rich in India.

We assume that the landlords and workers each have one unit of time which they must allocate between agricultural production and their alternative activities. The opportunity wage of a landlord is  $v$ , for a hired worker it is  $w$ , and for a labor-supervisor it is  $u$ , with  $u \geq w$ . The labor market is assumed to be competitive, and all opportunity incomes are assumed to be exogenously determined. The price of materials,  $p$ , is also assumed to be given.

The landlord has three options regarding the manner in which his land might be cultivated. First, he could self-cultivate by hiring unskilled labor and providing both management and supervision himself. This is the fixed wage contract. He could, instead, lease out the land to a tenant for a fixed lump sum rental. In this fixed rental contract, the tenant hires unskilled labor and provides both management and supervision himself. Finally, the landlord and the tenant could make a share contract in which the former provides management, the latter supervision, and output is shared. The share contract affords the opportunity for specialization—each agent performs the task at which he has the absolute advantage. In an arrangement in which management and supervision inputs are provided by different agents, the moral hazard of shirking arising from the unobservability of these inputs would be mitigated if both agents are made residual claimants. This is precisely what the share contract accomplishes. Thus sharecropping emerges in a natural fashion as a response to the possibility of achieving a superior input mix through resource pooling in the face of a moral hazard problem. Such an arrangement brings together superior inputs into production, but it suffers from the disadvantage of the disincentive effect arising from the fact that each of the two agents receives only a fraction of his marginal product.<sup>11</sup> These considerations, along with the opportunity incomes of the agents, will determine the contract that will emerge as the dominant one.

<sup>11</sup> This tradeoff has been pointed out by Bliss and Stern (p. 309).

We now turn to the determination of the optimal contract. For each type of contract we set up the optimization problem facing the agents.

#### A. Fixed Wage Contract

Assume the price of the agricultural output to be a constant,  $P$ . Under this type of contract, the landlord hires unskilled labor and allocates his time between management, supervision, and his alternative activity in order to maximize his expected net income:

$$(4) \quad \Pi_1^w = \max_{t_1, s_1, M, L} [Pf(t_1, \gamma_1 s_1, M, L, H) - pM - wL] + (1 - t_1 - s_1)v$$

$$0 \leq t_1 \leq 1, 0 \leq s_1 \leq 1, 0 \leq t_1 + s_1 \leq 1.$$

The term in the square brackets represents the landlords' expected income from cultivation and the other term the income from his alternative activity.

#### B. Fixed Rental Contract

Under this contract a worker leases the land for a fixed lump sum rental from the landlord. The expected net income of the tenant prior to paying the lump sum rental is given by

$$(5) \quad \Pi_2^r = \max_{t_2, s_2, M, L} [Pf(\gamma_2 t_2, s_2, M, L, H) - pM - wL] + (1 - t_2 - s_2)u$$

$$0 < t_2 \leq 1, 0 \leq s_2 \leq 1, 0 \leq t_2 + s_2 \leq 1.$$

We need to determine the fixed rent  $R$  that the landlord will demand of the tenant. Given the existence of a perfectly elastic supply of tenants, the rent on the land will be competed up until the tenant is at (or marginally above) his opportunity income  $u$ . Thus, we have

$$(6) \quad R = \max\{0, \Pi_2^r - u\}.$$

Under this type of contract, the landlord devotes all of his time to his alternative

activity. The landlord's income under the fixed rent contract is thus

$$(7) \quad \Pi_1^v = v + R.$$

### C. Share Contract

In the share arrangement, the landlord and the tenant each provide one of the unmarketed inputs and the profit (or output) is shared according to some endogenously determined, but mutually agreed upon, rule.

For the purpose of tractability, we make the assumption of complete specialization. It is easy to see that under this assumption the maximum expected output would be obtained when each agent provides the input in which he has the absolute advantage.

In what follows, it is convenient to define the restricted expected profit function,  $\Pi(t, s)$ , obtained by optimally choosing the amounts of labor and materials for parametrically given  $t$  and  $s$ :

$$(8) \quad \Pi(t, s) = \max_{M, L} Pf(t, s, M, L, H) - pM - wL.$$

We shall assume that the sharing rule takes the linear form:

$$(9a) \quad S_2 = \alpha + \beta \Pi,$$

where  $S_2$  is the expected profit share accruing to the tenant and  $\alpha, \beta$  are constants that are to be endogenously determined. The landlord's share of the expected profits must then be

$$(9b) \quad S_1 = -\alpha + (1 - \beta)\Pi.$$

The sharing rule (9) implies that what is shared is the revenue, net of labor and material costs, so that both agents bear part of the costs. This minor variation of the traditional assumption of output sharing is made purely for analytic convenience; the exercise can be repeated for an analogous output-sharing arrangement. In fact, it is typically the case that the share contract indicates not only the crop share but also

allocates the burden of costs. Implicit in (9) is the simplifying assumption that the labor and material costs are also shared in the same proportion as revenues. While this assumption can be relaxed, it would add little by way of insight to do so in what follows.

We now turn to the question of how the parameters  $\alpha$  and  $\beta$  of the share equations (9) are determined. The idea behind the procedure we use is simple. First, we examine the optimal decisions of the tenant and the landlord conditional on arbitrarily fixed values of  $\alpha$  and  $\beta$ . Then we step back and have the landlord choose  $\alpha$  and  $\beta$  in order to maximize his expected income subject to the tenant being at no less than his opportunity income. The landlord's choice is foresighted in that his decision is made under full knowledge of how the choice of  $\alpha$  and  $\beta$  will affect the ensuing equilibrium by affecting incentives.

Given the sharing rule (9), the tenant will choose the amount of time he will devote to supervision in order to maximize the variable component of his expected profits; he solves

$$(10) \quad \max_{s_2} \beta \Pi(t_1, s_2) + (1 - s_2)u, \quad 0 \leq s_2 \leq 1.$$

An analogous problem confronts the landlord; he must choose the amount of time to engage in management:

$$(11) \quad \max_{t_1} (1 - \beta) \Pi(t_1, s_2) + (1 - t_1)v, \quad 0 \leq t_1 \leq 1.$$

Following the literature, we shall assume that the two-person game we have here is resolved noncooperatively.<sup>12</sup> A natural assumption—and one which we invoke—is that each

<sup>12</sup> One could make a case that sharecropping involves cooperative rather than noncooperative behavior. If we assume the former, we would need to resort to bargaining theory to determine each agent's share. It is likely—as is true of the Nash bargaining solution—that the bargaining strength of each agent is determined by his payoff in the noncooperative resolution of the game being considered here.

agent entertains Nash conjectures, that is, he maximizes his expected profits taking as parametric the choice of the other agent.<sup>13</sup> The solution to (10) determines the Nash best response of the tenant:

$$(12a) \quad s_2 = \sigma(t_1; \beta),$$

while the solution to (11) yields the best-response function of the landlord:

$$(12b) \quad t_1 = \tau(s_2; \beta).$$

At a Nash equilibrium pair  $[t_1^*(\beta), s_2^*(\beta)]$ , equations (12a) and (12b) are simultaneously satisfied. The concavity of the production function implies that the restricted profit function (8) is concave in  $t$  and  $s$  (Theorem 2.20 in W. Erwin Diewert, 1973). This in turn, implies that for given  $\alpha$  and  $\beta$  the payoffs to the tenant and landlord (the objective functions of (10) and (11), respectively) are concave in  $t_1$  and  $s_2$ . Further, since the choice variables  $t_1$  and  $s_2$  are restricted to the compact convex set  $[0, 1]$  the existence of a Nash equilibrium is assured (see Theorem 7.4 in James Friedman, 1977). We shall assume that for given  $\alpha$  and  $\beta$ , the Nash equilibrium is unique.

The choices  $t_1$  and  $s_2$  in the Nash equilibrium will not, of course, depend on the parameter  $\alpha$ .<sup>14</sup> However, they will depend nontrivially on  $\beta$  since a change in this parameter, *ceteris paribus*, alters each agent's marginal remuneration and hence his incentives. For a given  $\beta$ , the landlord will set  $\alpha$  at a level that holds the tenant at (or barely above) his opportunity income. Thus  $\alpha$  satisfies

$$\alpha + \beta \Pi[t_1^*(\beta), s_2^*(\beta)] + [1 - s_2^*(\beta)]u = u,$$

so that

$$(13) \quad \alpha(\beta) = s_2^*(\beta)u - \beta \Pi[t_1^*(\beta), s_2^*(\beta)].$$

Finally, the landlord will choose  $\beta$  so as to maximize his expected income; the endogenous value,  $\beta^*$ , of  $\beta$  thus solves

$$(14) \quad \max_{\beta} -\alpha(\beta) + (1 - \beta) \Pi[t_1^*(\beta), s_2^*(\beta)] + [1 - t_1^*(\beta)]v.$$

We shall assume that  $t_1^*(\beta)$  and  $s_2^*(\beta)$  are continuous in their arguments. This ensures that a solution to the optimization problem (14) exists for  $\beta$  lying in the convex compact set  $[0, 1]$ . This completely endogenizes the parameters of the sharing rule (9). The landlord's expected income under the share contract is thus

$$\Pi_1^{sc} = -\alpha(\beta^*) + (1 - \beta^*) \times \Pi[t_1^*(\beta^*), s_2^*(\beta^*)] + [1 - t_1^*(\beta^*)]v.$$

Upon substituting for  $\alpha$  from (13), the above expression may be rewritten as

$$\Pi_1^{sc} = \Pi[t_1^*(\beta^*), s_2^*(\beta^*)] + [1 - t_1^*(\beta^*)]v + [1 - s_2^*(\beta^*)]u - u,$$

which, as one would expect, is the joint profit of the landlord and tenant, less the opportunity income of the latter.

Having determined the landlord's expected income under all three contractual arrangements, the determination of the arrangement that will prevail is, in principle, trivial: that contract would be chosen which maximizes the landlord's expected income. Since the supply of tenants is perfectly elastic, any rents that are generated accrue to the owner of the scarce resource, land: irrespective of the contract, the tenant earns only his opportunity income.

## II. Results

In this section we explicitly solve the model presented above for a Cobb-Douglas specification and present its comparative static results. The expected output of a farm, in the notation of the last section, is assumed to be

<sup>13</sup> It is interesting that substitution of the assumption of Nash behavior by Stackelberg behavior with the landlord as the leader results in the cooperative outcome.

<sup>14</sup> Unless, of course,  $\alpha$  is so negative that the tenant chooses not to participate at all in cultivation.



given by<sup>15</sup>

$$(15) \quad q = A t^{\delta_1} E^{\delta_2} M^{\delta_3} H^{\delta_4},$$

where  $A$ ,  $\delta_i$ ,  $i=1,2,3,4$  are positive constants and  $\sum_{i=1}^4 \delta_i = 1$ . The aggregator for the effort in (1) is also assumed to be Cobb-Douglas in structure:

$$(16) \quad E = s^\epsilon L^{1-\epsilon}, \quad 0 \leq \epsilon \leq 1.$$

The explicit solution of the model for Cobb-Douglas specifications (15) and (16) is presented in the Appendix. Since the restricted profit function (8) takes the form (A4) of the Appendix, which is concave in  $t_1$  and  $s_2$ , so are the objective functions (10) and (11) of the landlord and tenant, respectively, for any  $\beta \in [0,1]$  under sharecropping. Since  $t_1^*(\beta)$  and  $s_2^*(\beta)$  (given by A15a) and (A15b) of the Appendix), are continuous functions of  $\beta$ , so is the landlord's payoff (14). The solution to (14) determines the endogenous profit share of the tenant.<sup>16</sup> The determination of the fixed wage and fixed rental contracts are straightforward and are also presented in the Appendix.

We now present comparative static results, obtained by carrying out a series of parametric variations on the model. Parameter values in these exercises are chosen so as to facilitate a clear exposition of the forces that determine contractual structure. We note, however, that none of our conclusions—that are all qualitative in nature and are explained intuitively—are sensitive to the choice of parameter values. Three comparative static exercises are performed. We

analyze the impact of the following on the contractual structure:

(a) changes in the relative importance of the labor effort and management inputs (i.e., the effect of variations in  $(\delta_2/\delta_1)$ ;

(b) changes in the monitoring technology (i.e., the effect of variations in  $\epsilon$ );

(c) changes in the relative opportunity incomes of the two classes (i.e., effect of variations in  $(v/u)$ ).

A fourth exercise was conducted to study the variation of the optimal profit share,  $\beta^*$ , for a wide range in  $(\delta_2/\delta_1)$  and  $\epsilon$ .<sup>17</sup> The motivation for this exercise was to ask if our model would shed some light on the observed relative constancy of  $\beta$  in the neighborhood of one-half under a wide range of technological and market conditions.

The exogenous parameters of the model are those pertaining to the production function ( $A$ ,  $\epsilon$ ,  $\delta_i$ ,  $i=1, \dots, 4$ ), the distribution of abilities in management and supervision across classes ( $\gamma_1$  and  $\gamma_2$ )—henceforth referred to as the relative efficiency parameters—and the opportunity incomes ( $v$  and  $u$ ). In this entire section we normalize the price of output  $P$ , and the farm size  $H$  to unity. For a given configuration of parameter values, we numerically solve for the landlord's expected incomes under fixed wage and fixed rental contracts and compare the outcome with the landlord's expected income under the share contract.<sup>18</sup>

Figure 1 illustrates the partitioning of the relative efficiency (i.e.,  $\gamma_1$ ,  $\gamma_2$ ) parameter space according to the contractual arrangement that will prevail for parameter values noted in the figure. For low values of  $\gamma_1$  and  $\gamma_2$ , sharecropping prevails, since the diversification of activities (management and super-

<sup>15</sup>Note that (15) also lends itself to the interpretation of management as the monitoring of material input quality with  $t^{\delta_1} M^{\delta_3}$  as the aggregator that can be construed as the "effective material input."

<sup>16</sup>For any  $\beta \in [0,1]$  there are, in fact, two Nash equilibria, one of which is the trivial one  $t_1^*(\beta) = s_2^*(\beta) = 0$ ; since all inputs are essential, if either the landlord or the tenant applies zero effort, the Nash best response of the other is also to apply zero effort. Since this solution corresponds to the case in which there is no agricultural production at all, we shall restrict our attention in what follows to the nontrivial Nash outcome.

<sup>17</sup>For the chosen specification of the model, it turns out that  $\beta^*$  is invariant with respect to changes in opportunity incomes  $v$  and  $u$ . However, we would not expect this to be a feature of a more general specification.

<sup>18</sup>Note that under the fixed wage and fixed rental contracts, the entrepreneur has only one unit of time to allocate between management and supervision, while two units of time are available under a share contract. In order to be able to compare the outcomes of the various contracts without bias, we have chosen parameter values that yield interior solutions.

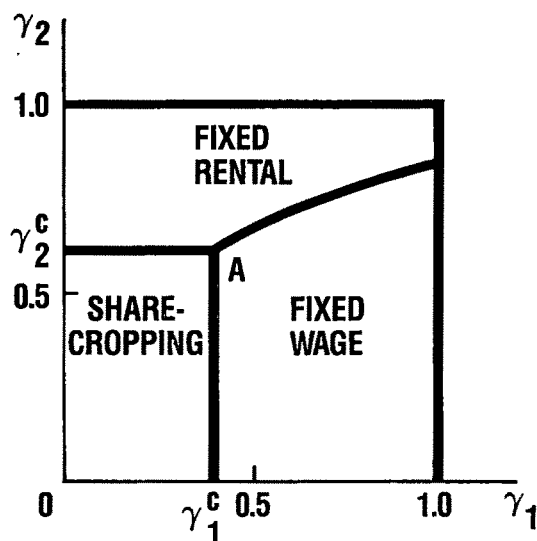


FIGURE 1. FRAGMENTATION OF RELATIVE EFFICIENCY PARAMETER SPACE ACCORDING TO DOMINANT CONTRACTURAL TYPE

(Parameter values:  $\delta_1 = 0$ ,  $\delta_2 = 0.2$ ,  $\delta_3 = \delta_4 = 0.25$ ,  $\varepsilon = 0.5$ ,  $A = 4$ ,  $p = w = 1$ ,  $u = 1.1$ ,  $v = 1.25$ .)

vision) required of the agents in the other two contractual arrangements render them relatively unprofitable. Such would be the case if the rural society is polarized into a landlord class alienated from day-to-day farm work and a working class devoid of managerial abilities. As markets develop, diffusing information and hence eliminating the disparity in managerial abilities between the two classes (i.e., as  $\gamma_2$  increases), sharecropping will give way to fixed rental contracts. To the extent that access to information is correlated with land ownership, a relatively uniform distribution of the latter would imply a high value of  $\gamma_2$ , rendering sharecropping less likely. In some states of India, where the distribution of land across the classes is highly skewed, sharecropping is commonly observed.

Sharecropping is often regarded as a backward or a feudal form of agriculture. This view is explained by Figure 1, according to which sharecropping would dominate when markets are either absent or underdeveloped and the class structure is polarized—two characteristics that are commonly associated with feudalism and backwardness.

If  $\gamma_2$  is small while  $\gamma_1$  is large, that is, if the workers lack managerial abilities and the landlords are not comparatively inept at supervision, wage contracts will obtain, as shown in Figure 1.

Absentee landlords typically appear to lease out their land under fixed rental contracts (Alston-Higgs; Bliss-Stern; Chao). This is quite consistent with our model: an absentee landlord typically does not provide any input at all apart from land and consequently sharecropping is not viable. Landowners who are engaged in some activity other than cultivation possess neither any supervisory abilities nor any advantage over the workers in access to information (i.e.,  $\gamma_1 \ll 1$  and  $\gamma_2 \approx 1$ ). The resulting contracts would therefore be of the fixed rental type.

Figure 1 may also be used to explain the rise of sharecropping in the postbellum South of the United States. With the abolition of slavery, the previously used supervision technology of the landlords was no longer feasible (i.e.,  $\gamma_1$  declined) and the absence of managerial ability among the emancipated slaves made sharecropping inevitable (Alston-Higgs; Reid, 1976). The lack of access of the freed men to credit and other agricultural inputs was probably also a contributing factor (Jaynes).

It is interesting to note that the ability to pool otherwise unobtainable resources under sharecropping allows the application of a superior input bundle to cultivation. This results in the possibility of the agricultural yield under sharecropping being higher than that under any other contract despite the disincentive effect operating in sharecropping. This may explain the observations reported by Bliss-Stern, and M. Mangahas et al. (1976), among others, that sometimes the yields under the share contract are higher than under alternative arrangements. In general, depending on the tradeoff between the advantage of a superior input combination and the disadvantage of the disincentive effect under sharecropping, the yield may compare favorably or unfavorably with other arrangements.

Figure 1 does not explain the coexistence of two or more types of contracts in the same geographical area. For ease and clarity we have cast our model in terms of only two

classes and one technology. It is easy to see that, in principle, the model could be extended to incorporate more than two classes, with each class differently endowed (i.e.,  $\gamma_1$  and  $\gamma_2$  are different across classes), and multiple technologies (for example, different crops). The introduction of this realistic feature of heterogeneity into the model would lead, in a natural fashion, to heterogeneity in the contractual structure in the same area at the same time. Thus, for example, absentee landlords may be found to lease out their land on a fixed rental basis to small farmers with some managerial ability but low opportunity incomes, while large farmers with access to market information might sharecrop with marginal farmers having none. At the same time, there can exist farmers who self-cultivate by hiring labor under wage contracts.

For what follows it is convenient to define the critical values,  $\gamma_1^c$  and  $\gamma_2^c$ , of the relative efficiency parameters  $\gamma_1$  and  $\gamma_2$ , respectively. The value  $\gamma_1^c$  ( $\gamma_2^c$ ) is that value of  $\gamma_1$  ( $\gamma_2$ ) at which, for given  $\gamma_2$  ( $\gamma_1$ ), the landlord would choose to switch from sharecropping to a fixed wage (fixed rental) contract, *ceteris paribus*. In other words,  $(\gamma_1^c, \gamma_2^c)$  denotes the coordinates of the point *A* in Figure 1. Much of the discussion to follow is cast in terms of  $\gamma_1^c$  and  $\gamma_2^c$ .

Figure 2 illustrates the result of the second comparative static exercise, viz, the effect of changes in  $(\delta_2/\delta_1)$ . An increase in the importance of management relative to labor effort (i.e., a decrease in  $(\delta_2/\delta_1)$ ) lowers  $\gamma_1^c$  and increases  $\gamma_2^c$ . In other words, the area covered by the fixed wage contract in Figure 1 increases, while that covered by the fixed rental contract shrinks. Precisely the opposite would be true if the importance of labor effort increases relative to management. This result, which is eminently in accord with intuition, explains the transformation of share contracts into wage contracts in India in response to the introduction of the High Yielding Variety (HYV) technology. The increased importance of know-how in production caused a decrease in  $(\delta_2/\delta_1)$  changing sharecropping to fixed wage contracts. By the same reasoning it also follows that if the initial situation were one predominantly of fixed rental contracts, a sufficiently large

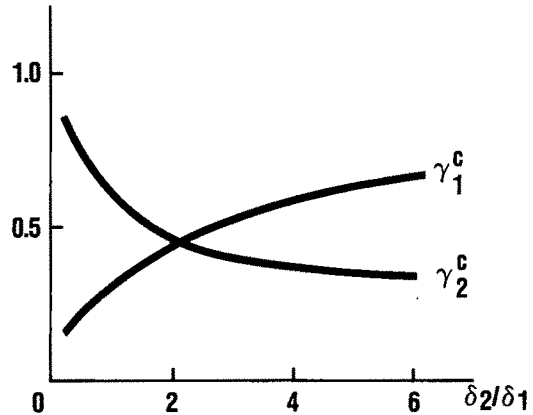


FIGURE 2. DEPENDENCE OF CRITICAL RELATIVE EFFICIENCY PARAMETERS ON  $(\delta_2/\delta_1)$

(Parameter values:  $\delta_1 + \delta_2 = 0.7$ ,  $\delta_3 = \delta_4 = 0.15$ ,  $\epsilon = 0.5$ ,  $A = 3$ ,  $p = w = 1.0$ ,  $u = v = 1.25$ .)

decline in  $(\delta_2/\delta_1)$  could result in a direct transition to a regime of fixed wage contracts, bypassing the share contract altogether. Over time, the diffusion of know-how reduces the advantage possessed by the landlord (i.e.,  $\gamma_2$  increases), and we might observe a reappearance of share contracts. Ranjit Sau (1976) reports some such cases in Haryana (India) ten to twelve years after HYV technology was first introduced in the area.

Figure 2 also explains an observation made by Reid (1979) that in the U.S. South there was a strong correlation between the extent of sharecropping and the labor intensity of production. Cotton, for example, which is a highly labor-intensive crop, was typically cultivated under sharecropping arrangements. The more labor intensive the crop, the greater the premium placed on supervision and, therefore, the less likely are the landlords to self-cultivate. This, together with the poor management abilities of the recently freed slaves (which made fixed rental tenancies unlikely), resulted in sharecropping as the dominant contractual arrangement.

Figure 3 demonstrates that  $\gamma_1^c$  increases and  $\gamma_2^c$  declines as the importance of supervision increases (i.e., as  $\epsilon$  increases). This is because as the activity in which the tenant has the advantage (supervision) acquires greater importance in the production process, some fixed wage contracts that previ-

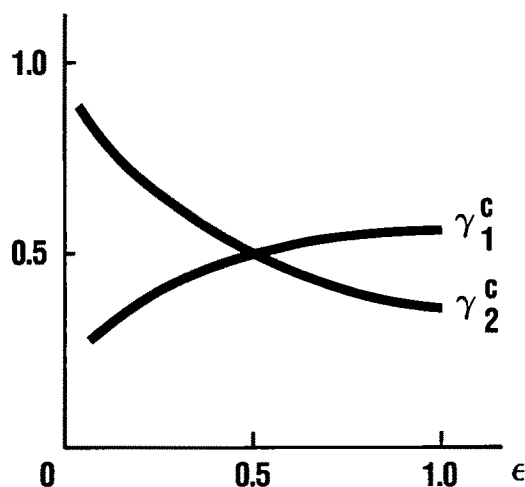


FIGURE 3. DEPENDENCE OF CRITICAL RELATIVE EFFICIENCY PARAMETERS ON  $\epsilon$

(Parameter values:  $\delta_1 = 0$ ,  $\delta_2 = 0.4$ ,  $\delta_3 = \delta_4 = 0.2$ ,  $A = 4$ ,  $p = w = 1$ ,  $u = v = 1.25$ .)

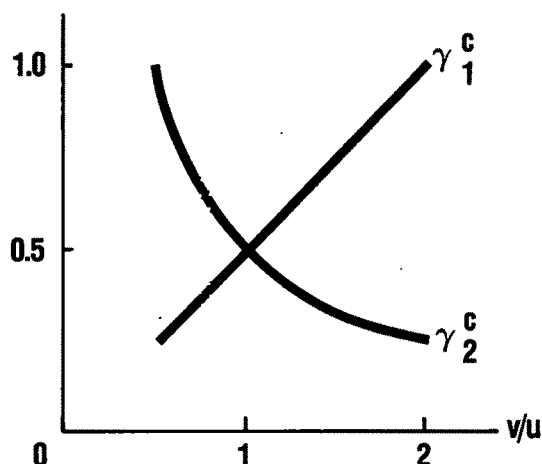


FIGURE 4. DEPENDENCE OF CRITICAL RELATIVE EFFICIENCY PARAMETERS ON RATIO OF OPPORTUNITY INCOMES

(Parameter values:  $\delta_1 = 0.2$ ,  $\delta_2 = 0.4$ ,  $\delta_3 = \delta_4 = 0.2$ ,  $\epsilon = 0.5$ ,  $A = 3$ ,  $p = w = 0.6$ ,  $u = 1.2$ .)

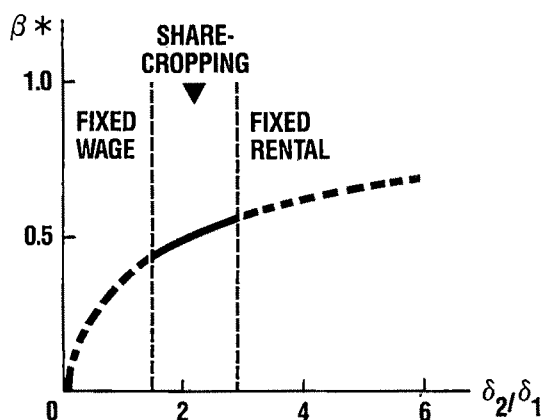
ously marginally dominated sharecropping would now give way to share contracts. Likewise, some share contracts that previously dominated fixed rental contracts would now give way to the latter.

The importance of labor monitoring depends on the variance of the quality of effort allowed by the production technology. Specialization of labor facilitates labor monitoring (i.e., reduces  $\epsilon$ ) through routinization and systematization of labor tasks. Often mechanization of activities implies that the pace and quality of output are primarily determined by machine specifications rather than by the quality of the operator's effort, thus reducing the importance of labor supervision (i.e., reducing  $\epsilon$ ); the machine supervises labor.<sup>19</sup> Thus, following a technological change involving either labor specialization or mechanization, we would expect to see a reduction in the supervision requirement (i.e., lower  $\epsilon$ ) and consequently a change from share and fixed rental contracts toward wage contracts. We thus have an additional explanation for the shift toward wage contracts

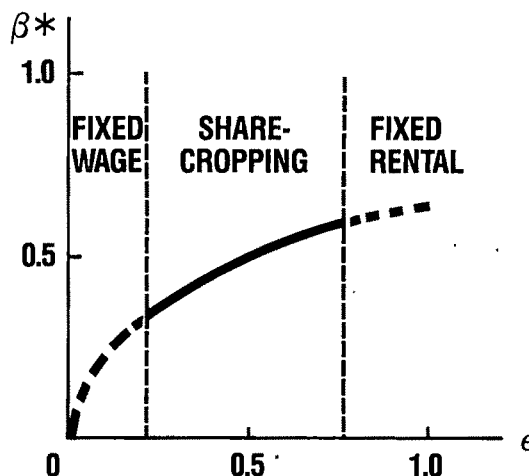
in Punjab (India) in the mid-1960's, since the introduction of the HYV technology was accompanied by mechanization. It is, therefore, important to consider the possible supervision consequences of labor displacing mechanization.

Figure 4 describes the impact of a change in the opportunity incomes on the contractual structure. An increase in the opportunity income of the landlord, *ceteris paribus*, decreases the area in  $(\gamma_1, \gamma_2)$  space corresponding to fixed wage contracts, while increasing that corresponding to fixed rental contracts (i.e.,  $\gamma_1^c$  increases and  $\gamma_2^c$  decreases). An implication of this result is that geographical areas in which the opportunity incomes of landlords are relative higher, *ceteris paribus*, are likely to exhibit a greater prevalence of fixed rental tenancy. Shigemochi Hirashima (1978, p. 56) has suggested that, in Pakistan, wage contracts became less prevalent as the supply of agricultural labor (as a percentage of total rural population engaged in cultivation) increased during the period 1901–31. This observation could be explained by the result presented in Figure 4; the increase in the percentage of agricultural laborers would *ceteris paribus* result in an increase in  $v/u$  and consequently an increase in  $\gamma_1^c$ .

<sup>19</sup>Alston (1978) offers convincing arguments as to why mechanization might reduce supervision requirements.

FIGURE 5. EFFECT OF  $(\delta_2/\delta_1)$  ON THE PROFIT SHARE

(Parameter values:  $\delta_1 + \delta_2 = 0.7$ ,  $\delta_3 = \delta_4 = 0.15$ ,  $\epsilon = 0.5$ ,  $A = p = w = 1$ ,  $u = v = 1.25$ ,  $\gamma_1 = \gamma_2 = 0.4$ )

FIGURE 6. EFFECT OF  $\epsilon$  ON THE PROFIT SHARE

(Parameter values:  $\delta_1 = 0.2$ ,  $\delta_2 = 0.4$ ,  $\delta_3 = \delta_4 = 0.2$ ,  $A = 4$ ,  $p = w = 1$ ,  $u = v = 1.25$ ,  $\gamma_1 = \gamma_2 = 0.4$ )

H. Laxminarayan and S. Tyagi (1977) while discussing the interstate variation in tenancy in India based on the NSS data (1970–71) and the Agricultural Census (1971–72) make an observation similar to the one made by Hirashima.

Suppose that the industrialization strategy pursued by a developing country is such that the demand for people with greater human capital endowment is created at a faster rate than for unskilled workers, without a corresponding change in the supply. To the extent that the members of landlords' families are likely to have had a greater opportunity to acquire human capital, a consequence of such industrialization would be an increase in  $v$ , but not in  $u$ . Figure 4 suggests that in the face of a steady increase in  $(v/u)$ , we would expect to see a transition from fixed wage contracts to sharecropping; a further increase in  $(v/u)$  would result in a transition from sharecropping to fixed rental contracts with absentee landownership. On the other hand, in instances where  $(v/u)$  decreases, there would be a shift towards wage contracts. Taiwan may be cited as one such case.

A persistent puzzle in the sharecropping literature is why the share,  $\beta$ , is often observed to be in the vicinity of one-half; it rarely, if ever, takes on values at either extreme of its range. Figures 5 and 6 offer a possible explanation. Figure 5 depicts the endogenized value of  $\beta$  when  $(\delta_2/\delta_1)$  is

varied keeping all else constant, and Figure 6 shows the endogenized  $\beta$  as a function of  $\epsilon$ . The entire range of  $\beta^*$  indicated in the figures will, however, not necessarily be observed when  $(\delta_1/\delta_2)$  and  $\epsilon$  are varied. When we also allow for fixed wage and fixed rental contracts, sharecropping may not be the dominant contract. The shares in the range where the share contract is the dominant one are represented by the solid portions of the curves in Figures 5 and 6; the dashed portions represent situations where sharecropping is dominated, although  $\beta$  is optimally chosen. Thus only the solid segments of these curves represent *realizable* shares, since outside this range sharecropping will not be observed.

From these figures we see that the variation in the realizable (and thus observable) values of  $\beta^*$  around one-half is minimal, despite the relatively large polarization assumed in the abilities of the landlords and tenants ( $\gamma_1 = \gamma_2 = 0.4$ ). While the bounds on the observable values of  $\beta$  depend on the specific parameter values chosen, these results shed some light on why the share is found to be in the vicinity of one-half. A large deviation of the share from one-half would exacerbate the disincentive effect by diluting incentives for either the tenant or the landlord, which would result in share-

cropping being dominated by either the fixed wage or the fixed rental contract. Thus when sharecropping is the observable outcome, the share is neither inordinately high nor inordinately low.<sup>20</sup> We are in agreement with Bliss and Stern in that the choice of the share is probably rooted in tradition and one-half has the appeal of equity. Any attempt to explain why it must be exactly one-half is bound to be contrived, besides being empirically false. What the above results indicate is that middling values of the share are consistent not only with equity considerations, but also with efficiency requirements.

### III. Summary and Conclusions

In this paper we have formally modeled the idea that each contractual form entails a different type of agent (landlord or tenant) providing unmarketed factor inputs. Sharecropping is viewed as a partnership in which each partner provides the unmarketed factor input in which he is the better endowed. The contractual structure has been endogenized by solving the problem of optimal time allocation in cultivation by each agent in terms of exogenous parameters such as the opportunity incomes, characteristics of the prevailing technology, and attributes of the economic agents. The model has then been applied to explain some variations over time in the contractual structure in India and the United States.

Our view that sharecropping is a partnership is consistent with three significant empirical observations made in the literature. First, the yields on farms cultivated under sharecropping are sometimes found to be higher than on farms alternatively cultivated, despite the moral hazard inherent in the noncooperative nature of the share contract. In our model this may result from the advantage of being able to pool unmarketed resources. Second, the hypothesis that the greater the production risk, the greater the

prevalence of sharecropping, has received little support in recent empirical investigations (Rao, 1971; Pant, 1981). These studies cast some doubt on the alternative hypothesis that risk sharing is the principal motivation behind sharecropping. Third, the relative insensitivity of the share  $\beta$  to the variation in technology and market characteristics across different regions has also been shown to be consistent with our model.

It should be pointed out that although "management" and "supervision" are the two unmarketed resources relevant to the specific cases we have tried to explain, the model could be cast in terms of *any* unmarketed resources. What is crucial to the model is that there be no markets in which the contract-determining resources can be traded independently of the time of the resource owner. We do not analyze in this paper the interesting question as to why markets might not exist for certain resources.

Our model implies that in sharecropping arrangements, landlords contribute some unmarketed resource, while in fixed rental arrangements they do not. In reality, one would expect some landlord participation—though to a smaller extent—even in fixed rental arrangements due to long-run considerations (such as the maintenance of soil productivity, etc.). Reid (1979, p. 295) has presented evidence that, in the U.S. post-bellum South, landlord participation was greater in share contracts than in fixed rent contracts. Bliss and Stern (p. 127) and Thomas Smith (1959, pp. 153–55) present similar evidence on contracts in India and Japan, respectively. We have demonstrated that the implications of our theory are consistent with observations drawn from a diverse set of countries and time periods. The prevalence of the metayage system in continental Europe in the eighteenth and nineteenth centuries and of share contracts in the present day U.S. Midwest may, however, prove to be anomalies for our theory.

The assumption of only two classes and a single production function inevitably leads to a single dominant contract. The model can, however, be extended to incorporate a heterogeneous rural population with more than two classes or more than one produc-

<sup>20</sup> It should be noted that although we have cast the problem in terms of profit shares, the observations made in the literature usually refer to crop shares. This intuition is, however, equally valid for crop shares.

tion function (for example, different crops in the same area), or both. It is easy to see that such an extension could explain the coexistence of different contracts in the same area and would also be consistent with "the agricultural ladder" hypothesis.<sup>21</sup>

Our model assumes that tenants are landless. This assumption could be dropped at the cost of a minor increase in computational difficulty. Such a modification would allow us to explain the puzzling reversal of the traditional picture of a poor tenant and a rich landowner. It has been observed that sometimes rich farmers armed with managerial ability and farm machinery (i.e., low  $\epsilon$ , and low  $\delta_2/\delta_1$ ) might want to lease land from poor landowners at a fixed rental (Nripen Bandyopadhyay, 1975; M. Nadkarni, 1976). The fixed rental tenants engaged in tobacco cultivation cited by Rao (1971) also belong to this category. The motivation, of course, is to adjust the operational holdings to match their high resource endowment.

We have followed the literature in modeling sharecropping as a one-period noncooperative game. We believe, however, that typically tenancy relationships are of a long-run nature. In addition, there are many implicit instruments (for example, consumption credit) through which the landlord may exercise control over the tenant to approximate the cooperative outcome. The exercise we have performed in this paper can be repeated under the assumption of a cooperative game.

The analysis of institutional arrangements presented here may have interesting implications for the study of similar arrangements in industrialized countries. The institutions of sharecropping and fixed rental tenancy resemble, for example, the revenue-sharing and royalty-charging franchise arrangements common in the American service sector as suggested by Peter Murrell. Similarly the technology transfers across national boundaries involve tenancy-like arrangements, for example, entirely owned subsidiaries by multinationals (wage contracts), joint ventures between multinationals and domestic

firms (share contracts), and royalty-charging arrangements (fixed rentals). Analysis similar to that employed in this paper could be used to study these arrangements.

#### APPENDIX

For the Cobb-Douglas specifications (15) and (16) the function  $f$  in (3) of a farm takes the form:

$$(A1) \quad f(t, s, M, L, H; \epsilon) = At^{a_1}M^{a_2}s^{a_3}L^{a_4}H^{a_5},$$

where

$$(A2) \quad a_1 = \delta_1, \quad a_2 = \delta_2, \quad a_3 = \epsilon\delta_3, \\ a_4 = (1 - \epsilon)\delta_3, \quad a_5 = \delta_4,$$

with

$$(A3) \quad \sum_{i=1}^5 a_i = 1.$$

A straightforward maximization reduces the restricted profit function (8) to the form

$$(A4) \quad \Pi(t, s) = Dt^{b_1}s^{b_2},$$

where

$$(A5) \quad b_1 = a_1/c, \quad b_2 = a_3/c, \quad c = 1 - a_2 - a_4,$$

$$(A6) \quad D = PAH^{a_5}\bar{M}^{a_2}\bar{L}^{a_4} - p\bar{M} - w\bar{L},$$

with

$$(A7a) \quad \bar{M} = \left[ PAH^{a_5} \frac{a_2}{p} \left( \frac{a_4 p}{a_2 w} \right)^{a_4} \right]^{1/c},$$

$$(A7b) \quad \bar{L} = \left[ PAH^{a_5} \frac{a_4}{w} \left( \frac{a_2 w}{a_4 p} \right)^{a_2} \right]^{1/c}.$$

1. *Fixed Wage Contract.* The landlord's expected profit  $\Pi_1^w$  is

$$(A8) \quad \max_{t_1, s_1} D^w t_1^{b_1} s_1^{b_2} + v(1 - t_1 - s_1)$$

$$0 \leq t_1 \leq 1, 0 \leq s_1 \leq 1, 0 \leq t_1 + s_1 \leq 1,$$

<sup>21</sup>For a summary description of the hypothesis, see Reid (1977, p. 405; and 1979, p. 300).

where  $D^{fw}$  is obtained from  $D$  by replacing  $A$  by  $A\gamma_1^{a_3}$  in (A7a), (A7b), and (A6).

Defining  $d = 1 - b_1 - b_2$ , the solution to (A8) is given by

$$(A9a) \quad t_1 = \left[ D^{fw} \frac{b_1}{v} \left( \frac{b_2}{b_1} \right)^{b_2} \right]^{1/d}; \quad s_1 = \frac{b_2}{b_1} t_1,$$

when  $t_1 + s_1 < 1$ , and

$$(A9b) \quad t_1 = \frac{b_1}{b_1 + b_2}; \quad s_1 = \frac{b_2}{b_1 + b_2}$$

otherwise.

2. *Fixed Rental Contract.* The tenant's expected profit,  $\Pi_2^f$ , prior to paying the rental is given by

$$(A10) \quad \max_{t_2, s_2} D^{fr} t_2^{b_1} s_2^{b_2} + u(1 - t_2 - s_2)$$

$$0 \leq t_2 \leq 1, 0 \leq s_2 \leq 1, 0 \leq t_2 + s_2 \leq 1,$$

where  $D^{fr}$  is obtained from  $D$  by replacing  $A$  by  $A\gamma_1^{a_1}$  in (A7a), (A7b), and (A6).

The solution to (A10) is given by

$$(A11) \quad t_2 = \left[ D^{fr} \frac{b_1}{v} \left( \frac{b_2}{b_1} \right)^{b_2} \right]^{1/d}; \quad s_2 = \frac{b_2}{b_1} t_2,$$

when  $t_2 + s_2 < 1$ , and (A9b) holds.

The landlord's expected payoff is then

$$(A12) \quad \Pi_1^f = \max[\Pi_2^f - u, 0] + v.$$

3. *Share Contract.* The reaction function of the tenant, which is the solution to (10), is given by

$$(A13) \quad s_2 = \min \left\{ 1, \left[ \frac{b_2 \beta D}{u} t_1^{b_1} \right]^{1/(1-b_2)} \right\}.$$

The landlord's reaction function, which is the solution to (11), is given by

$$(A14) \quad t_1 = \min \left\{ 1, \left[ \frac{b_1(1-\beta)Ds_2^{b_2}}{v} \right]^{1/(1-b_1)} \right\}.$$

Assuming an interior solution, the Nash equilibrium for given  $\beta$  is obtained as

$$(A15a) \quad t_1^*(\beta) = \left[ \frac{b_1 D}{v} \left( \frac{b_2 v}{b_1 u} \right)^{b_2} \right]^{1/d}$$

$$\times [\beta^{b_2} (1-\beta)^{1-b_2}]^{1/d},$$

$$(A15b) \quad s_2^*(\beta) = \left[ \frac{b_1 D}{v} \left( \frac{b_2 v}{b_1 u} \right)^{1-b_1} \right]^{1/d}$$

$$\times [\beta^{1-b_1} (1-\beta)^{b_1}]^{1/d}.$$

Finally, the landlord chooses  $\beta$ , as in (12), by solving

$$(A17) \quad \max_{\beta} D t_1^{*b_1}(\beta) s_2^{*b_2}(\beta)$$

$$+ v[1 - t_1^*(\beta)] - u s_2^*(\beta).$$

## REFERENCES

- Allen, Franklin, "On Share Contracts and Screening," *Bell Journal of Economics*, Autumn 1982, 13, 541-47.
- Alston, Lee J., "Costs of Contracting and the Decline of Tenancy in the South, 1930-1960," unpublished doctoral dissertation, University of Washington, 1978.
- and Higgs, Robert, "Contractual Mix in Southern Agriculture since the Civil War: Facts, Hypotheses and Tests," *Journal of Economic History*, June 1982, 42, 327-55.
- Bandyopadhyay, Nripen, "Changing Forms of Agricultural Enterprise in West Bengal," *Economic and Political Weekly*, April 26, 1975, 10, 700-01.
- Bardhan, Kalpana, "Rural Employment, Wages and Labour Markets in India, A Survey of Research—II," *Economic and Political Weekly*, July 2, 1977, 12, 1062-74.
- Bell, Clive and Zusman, Pinhas, "New Approaches to the Theory of Rental Contracts in Agriculture," mimeo., Development Research Centre, World Bank, August 1979.
- Binswanger, Hans P. and Rosenzweig, Mark R., "Contractual Arrangements, Employment and Wages in Rural Labor Markets: A



- Critical Review," in their *Rural Labor Markets in Asia: Contractual Arrangements, Employment and Wages*, New Haven: Yale University Press, 1982.
- Bliss, Christopher J. and Stern, Nicholas H., *Palanpur: The Economy of an Indian Village*, Oxford: Clarendon Press, 1982.
- Chao, K., "Tenure Systems in Traditional China," *Economic Development and Cultural Change*, January 1983, 31, 295-314.
- Cheung, Steven N.S., *The Theory of Share Tenancy*, Chicago: University of Chicago Press, 1969.
- Day, Richard H., "The Economics of Technological Change and the Demise of the Sharecropper," *American Economic Review*, June 1967, 57, 427-49.
- Diewert, W. Erwin, "Functional Forms for Profit and Transformation Functions," *Journal of Economic Theory*, June 1973, 6, 284-316.
- Friedman, James W., *Oligopoly and the Theory of Games*, Amsterdam: North-Holland, 1977.
- Hallagan, William, "Self-Selection by Contractual Choice and the theory of Sharecropping," *Bell Journal of Economics*, Autumn 1978, 9, 344-54.
- Hirashima, Shigemochi, *The Structure of Disparity in Developing Agriculture*, Tokyo: Institute of Developing Economies, 1978.
- Shikawa, Shigeru, *Essays on Technology, Employment and Institutions in Economic Development*, Tokyo: Institute of Economic Research, Hitotsubashi University, 1981.
- Jaynes, Gerald David, "Economic Theory and Land Tenure," in H. Binswanger, and M. Rosenzweig, eds., *Rural Labor Markets in Asia: Contractual Arrangements, Employment and Wages*, New Haven: Yale University Press, 1982.
- Laxminarayan, H. and Tyagi, S., "Tenancy, Extent and Inter-State Variations," *Economic and Political Weekly*, May 28, 1977, 12, 880-83.
- Mangahas, M., Miralao, V. and de los Reyes, R., *Tenants, Lessees, Owners, Welfare Implications of Tenure Change*, Quezon City: Ateneo de Manila University Press, 1976.
- Murrell, Peter, "The Economics of Sharing: A Transactional Cost Analysis of Contractual Choice in Farming," *Bell Journal of Economics*, Spring 1983, 14, 283-93.
- Nadkarni, M., "Tenants from the Dominant Class: A Developing Contradiction in Land Reforms," *Economic and Political Weekly*, December 25, 1976, 11, A137-55.
- Newbery, David M. G. and Stiglitz, Joseph E., "Sharecropping, Risk Sharing and the Importance of Imperfect Information," in J. Roumasset et al., eds., *Risk, Uncertainty and Agricultural Development*, New York: Agricultural Development Council, 1979.
- Pant, Chandrashekar, "Tenancy in Semi-Arid Tropical Villages of South India: Determinants and Effects on Cropping Patterns and Input Use," ICRISAT Progress Report, No. 20, May 1981.
- \_\_\_\_\_, "Tenancy and Family Resources: A Model and Some Empirical Analysis," *Journal of Development Economics*, February 1983, 12, 27-40.
- Rao, C. H. Hanumantha, "Uncertainty, Entrepreneurship and Sharecropping in India," *Journal of Political Economy*, June 1971, 79, 578-95.
- \_\_\_\_\_, *Technological Change and Distribution of Gains in Indian Agriculture*, Delhi: Macmillan Company of India, 1977.
- Reid, Joseph D., Jr., "Sharecropping as an Understandable Market Response: The Post-Bellum South," *Journal of Economic History*, March 1973, 33, 106-30.
- \_\_\_\_\_, "Sharecropping and Agricultural Uncertainty," *Economic Development and Cultural Change*, April 1976, 24, 549-76.
- \_\_\_\_\_, "The Theory of Share Tenancy Revisited—Again," *Journal of Political Economy*, April 1977, 403-07.
- \_\_\_\_\_, "Sharecropping in American History," in J. Roumasset et al., eds., *Risk, Uncertainty and Agricultural Development*, New York: Agricultural Development Council, 1979.
- Sau, Ranjit, "Can Capitalism Develop in Indian Agriculture?," *Economic and Political Weekly*, December 25, 1976, 11, A126-36.
- Smith, Thomas C., *The Agrarian Origins of Modern Japan*, Stanford: Stanford University Press, 1959.
- Stiglitz, Joseph E., "Incentives and Risk Sharing in Sharecropping," *Review of Economic Studies*, April 1974, 41, 219-55.

# Oil Field Unitization: Contractual Failure in the Presence of Imperfect Information

By STEVEN N. WIGGINS AND GARY D. LIBECAP\*

Private contracting is a solution to problems of production and exchange when transactions costs are low, but if they are high, contracting may be less successful (Ronald Coase, 1960; Oliver Williamson, 1976; Victor Goldberg, 1976). Accordingly, incorporation of transactions costs into economic analysis is necessary to determine when private contracting will be effective and when it will not. There has been, however, little empirical analysis, based upon testable theories, of the impact of transactions costs on contracting. This paper presents such an analysis, and shows that imperfect information can seriously limit the effectiveness of private contracting. The case considered is the widespread failure of private crude oil producing firms to unitize U.S. oil fields to reduce rent dissipation. Rent dissipation follows as multiple firms compete for migratory oil in common oil pools. Competitive production leads to excessive wells and surface storage, higher extraction costs because subsurface pressures are inefficiently depleted, and reduced overall oil recovery. Unitization is the obvious private contractual solution to rent dissipation. Under

unitization a single firm is selected to develop the reservoir with net returns shared by all parties, including firms that would otherwise be producing. As early as 1916, the U.S. Bureau of Mines called for unitization of U.S. oil fields; yet, by 1947, Joe Bain found only 12 fully unitized fields out of some 3,000 U.S. fields sampled (p. 29).<sup>1</sup> Our forthcoming paper (1985) shows that as late as 1975 neither Oklahoma nor Texas, two leading producing states, had as much as 40 percent of production from field-wide units.

We argue that the principal causes of contractual failure are imperfect and asymmetric information that prevent agreement on lease values and hold-out strategies of firms to increase their share of unit rents.<sup>2</sup> This study is based on an empirical analysis of unitization contracting in seven oil fields. Our data are from trade journals and company records, including detailed engineering studies, estimates of parameters affecting lease values, unit share allocation formulas, and votes on shares. Besides these quantitative data, the company files include minutes of negotiations on bargaining strategy and let-

\*Department of Economics, Texas A&M University, College Station, TX 77843, and Department of Economics and Karl Eller Center, University of Arizona, Tucson, AZ 85721, respectively. We thank Ray Battalio, Victor Goldberg, Jim Griffin, Jan Kmenta, James Smith, Oliver Williamson, and the referees for helpful comments. We also thank participants in university seminars at Arizona, Duke, Houston, the Saarlands, Texas A&M, Washington, and Yale. Capable research assistance was provided by Phil Mizzi. Funding from the NSF under grant SES-8207826 and the Center for Energy and Mineral Resources at Texas A&M is gratefully acknowledged. The order of names is random. Finally, we acknowledge our debt to the late Bruce A. Landis, Jr., a widely recognized unitization expert, whose aid in our gaining access to negotiation records made this research possible.

<sup>1</sup>In 1914, the U.S. Bureau of Mines estimated annual losses from competitive extraction at \$50 million, approximately one-quarter of the total value of U.S. production. In 1924, the Federal Oil Conservation Board was organized to review wasteful oil extraction, and it specifically endorsed unitization.

<sup>2</sup>Stephen McDonald (1979, p. 137) and others have pointed to hold-out strategies by firms for increased shares of unitized rents which block agreement. Our analysis reveals that this explanation is incomplete. A simple hold-out strategy could be employed by any firm, and it would be unrelated to structural advantage or the stage of field development. Yet, we find there are systematic differences, based upon structural conditions, across firms in the willingness of firms to join units. Further, we find that unitization contracts can be successfully completed late, as fields near depletion, or very early before production heterogeneities among firms are known.

ters regarding votes, lease parameters, and estimated gains from agreement.<sup>3</sup> From this unusually rich data source we construct a detailed analysis of negotiations on the seven fields and isolate the causes of contractual failure. Section I outlines the general contracting problem. Section II develops a simple theory of contracting in the presence of imperfect information, and Section III provides the empirical analysis.

## I. The Contracting Problem

### A. Contracting in General

Coase, Goldberg, Williamson, and others have argued that transactions costs can seriously disrupt private contracting. Goldberg and Williamson, for example, explore cases where *ex post* information asymmetries cause contracting to fail when complete contingent claims contracts cannot be written. Contingent updates are difficult to incorporate when new information does not arrive symmetrically across agents. Coase examines a different case where large numbers of parties prevent contracts from being written. In this paper we explore a third and potentially more important cause of contractual failure, *ex ante* information problems that occur during contract negotiation. In unitization contracting there are only a small number of firms involved, typically fewer than 15, so that large numbers problems frequently associated with transactions costs are not encountered. Further, firms do not rely on later events to reveal more accurate information to update contracts. Instead, the unitization bargaining problem occurs before contract execution and centers on the establishment of once-and-for-all unit shares. Dispute focuses on differences in lease value estimates, since lease values determine unit shares. Differences in value estimates arise because of general uncertainty regarding res-

ervoir dynamics and information asymmetries among the bargaining parties. Those disputes block consensus on unit shares, and contribute to contractual failure.

The structural characteristics that lead to contractual failure in unitization are also present in many other contracting contexts. In unitization, bargaining disputes regarding the relative values of leases to be included in the unit lead to a breakdown of negotiations for a sharing formula. It is clear that these same factors are common in other settings, where it is asserted that contracting will provide an efficient, low-cost solution to problems of allocation and exchange. A conceptually similar problem is contracting between firms and workers in team production, when monitoring is imperfect (Richard Startz, 1983). Another related case is private contracting between consumers and firms to reduce pollution. In both cases it is difficult to determine the *ex ante* contribution of each party to production, and, if successful, the contract changes the nature of production so that *ex post* adjustments are not generally feasible. These cases suggest that empirically untested optimism regarding the power of private contracting in the presence of imperfect information may be unwarranted and that more systematic analysis is needed.

### B. Contracting for Unitization

Unitization raises field rents by increasing oil recovery and reducing production costs compared to common pool production. Common pool conditions arise as multiple firms extract oil from the same reservoir. Land owners grant firms access to the reservoir through leases, and fragmented land ownership results in numerous firms exploiting the pool. Each is motivated to competitively drill and extract oil. Within the reservoir, oil is migratory, and property rights to it are assigned only upon extraction. Rapid production by a firm lowers subsurface pressure around its wells, stimulating oil immigration, which increases its share of total output. In the aggregate, these production strategies raise marginal extraction costs and reduce total recovery. High extraction rates deplete natural gas throughout the field, making oil

<sup>3</sup>The company records are from one of the largest producing firms in the United States. We were granted access to their unitization files. For reasons of confidentiality we cannot reveal the names of the bargaining firms on the seven fields examined in the paper.

more viscous and forcing costly artificial lift and injection of natural gas and water to raise pressure. Further, as natural gas leaves solution, pockets of oil become permanently trapped. With unitization these problems can be mitigated by having only a single firm develop the field with net revenues shared by all firms.

The division of net revenues is the central issue in unitization negotiations. The sharing formula is specified at unit agreement, and it assigns *once-and-for-all* shares of unit rents. A permanent share assignment is required because reservoir dynamics and relative lease production potential are fundamentally altered under unitization. Under unit management some wells are plugged and others are converted to gas injection to maintain pressure. No direct production occurs from these leases, and the oil originally below them is extracted elsewhere. Further, widespread injection of natural gas and water increases pressure in certain parts of the field, altering oil migration patterns. Because of these technical changes in the dynamics of production, unit shares must be based on *pre-unitization* estimates of lease values, and share negotiations cannot be reopened later to adjust for new contingencies.

Unit shares are based on estimates of lease values, but general uncertainty and asymmetric information block consensus on value estimates and, hence, shares. The parameters influencing lease values and unit shares include current and cumulative oil and gas production, number of wells, surface acreage, bottom hole pressure, gross acre feet of pay (volume of the producing formation), net acre feet of pay (nonporous and non-oil-bearing rock is subtracted from the gross measure), and remaining reserves (original oil-in-place less cumulative production). The first four parameters are directly observable and uncontroversial. The last four parameters describe reservoir characteristics under each lease, and provide more complete information on lease potentials. These are estimated from well logs and production histories, and require highly subjective interpretation by geologists and engineers. The need for subjective interpretation using arbitrary procedures leads to serious disputes. For ex-

ample, it was noted during negotiations in the Western RKM unit in Texas that: "The Engineering Committee could not agree upon oil reserves for a large number of tracts in the unit area because of the poor quality and interpretive nature of the available basic data" (Letter, Western RKM Unit File, Company Records).

The information problems that limit the number of parameters that can be used in allocation formulas is further illustrated in the calculation of remaining primary oil reserves. This parameter is a principal component of nonunitized lease value because it is an estimate of oil that may be produced from the lease under competitive production. It establishes a benchmark against which the value offered under unitization can be evaluated. Yet, remaining primary reserves are estimated using simple ordinary least square (*OLS*) regressions on specific functional forms that are often inaccurate. For example, in unsuccessful unit negotiations on the Wasson field in 1971, ultimate primary recovery (cumulative production plus remaining primary reserves) was estimated at 48 million barrels. This was based on production decline curves inferred from a production history of 36 million barrels; thus, it was estimated that one-quarter of the field's primary reserves remained. In 1978 negotiations were reopened after 2 million additional barrels had been produced. Ultimate primary was reestimated at 43 million barrels. A 6 percent change in output led to new information and a revision of remaining reserves estimates by approximately 50 percent (Wasson Unit File, Company Records).

Because of the subjective nature and wide variation in estimates of subsurface parameters, negotiating parties rely upon a small set of objectively measurable variables, but they are likely to be poor indicators of lease value. The problem is illustrated in Table 1 with regression estimates of output per acre using objective parameters for leases on three fields. Output per acre is a key determinant of lease value. In all three fields there remains a large, unexplained residual variance in the estimates due to inherent variation in reservoir quality and differences in the stage of development among leases that are not re-

TABLE 1—REGRESSIONS OF OUTPUT PER ACRE<sup>a</sup>

	Independent Variables:			Statistics		
	Intercept	Cumulative Output/Acre	Wells/Acre	Mean	SER	Coefficient
Goldsmith/Landreth	28.23 (9.79)	.036 (.0063)	38.87 (196.46)	68	46	68
North Cowden	40.19 (35.19)	.0083 (.0012)	618.69 (1471.58)	85	33	39
Prentice Northeast	2.45 (11.15)	.045 (.0062)	2602.88 (911.15)	116	39	34

<sup>a</sup>Dependent variable: Output per acre; standard errors are shown in parentheses.

flected in measurable parameters. For example, mean output per acre on Goldsmith/Landreth is 68 barrels with a standard error of the regression (*SER*) of 46 barrels. Similarly on North Cowden mean output is 85 barrels with *SER* 33, and on Prentice mean output is 116 barrels with *SER* 39. With so much unexplained variation in lease productivity there are ample grounds for disagreement about how long production differences will persist and whether they will grow larger or smaller over time. It is important to note that in such disputes the subjective judgment of a firm's engineers, who are familiar with individual well performance, provides an important source of *private* information about lease values. These judgements, however, are not easily verified by other parties, and accordingly, are difficult to incorporate into the unit allocation formula.

The problems of estimation and extrapolation for static structural characteristics are compounded in the estimation of dynamic reservoir performance characteristics, which are central to the negotiation of relative unit shares. Efforts to predict future performance are typically *ad hoc*. For example, predicted future production is often a simple extrapolation of past production with little systematic account taken of the pattern of water encroachment or other key variables. Such predictions are sensitive to the specific functional forms chosen. Companies often have differing, and strong, opinions about the correct estimation procedure, when choices may reallocate millions of dollars, and there is no generally accepted standard. The result is that consensus cannot be achieved on future

lease output. Such an agreement, however, is necessary for successful unit share negotiations.

These problems are particularly important for highly productive leases that are constrained by prorationing output quotas under state regulation. To estimate the well's future producing capability, it is necessary to have observations of the rate of production decline, which varies substantially across leases. On highly productive leases, producing the maximum regulated allowable, however, such observations are not available. Hence, the data at which the decline will begin and its rate must be based on the subjective opinions of engineers and geologists, which are open to dispute. Below we review cases where such problems impede unit agreements because parties cannot agree on potential future production. These information problems are central to the failure of private contracting.<sup>4</sup> We now develop a formal theory for analyzing unitization contracting.

<sup>4</sup>The unitization contracting problem is summarized by Raymond Myers:

The principal obstacle to full, voluntary agreement is the problem of dividing the proceeds of production. If development of the area sought to be unitized is complete, ... some lessors and leases may be inclined to rely on the possibility that... the entire production from their land will be more valuable than an undivided interest in production from a much larger unit. If development of the pool is relatively complete, there is frequent acrimony as to the respective shares of production to be given owners of interest in favorable parts of the structure and owners of interests in less favorable areas.... [1967, p. 108]

## II. Theory: Contracting for Unitization Agreements

This section presents a simple, testable model of a firm's decision to join a unit. The decision is made so as to maximize the expected present value of the firm's lease(s). There are two central features of assessing the relative values of oil leases that lead to contractual failure. One feature is that each firm has access to better information concerning the highly uncertain value of its own leases than do other firms in the field. It is assumed that at least a portion of this private information cannot be successfully communicated.<sup>5</sup> A second key aspect is that firms differ regarding the transformation of raw data into value estimates, and there is no absolute standard that can resolve differences. Either of these information problems is sufficient to cause the observed contractual breakdown, and both appear to be empirically important.

Let the raw, publicly available information concerning lease  $i$  be a vector of random variables  $x_i$  and the privately available information be random variables  $z_i$ . Firms have the incentives to develop procedures to share information. As discussed above, however, it is unlikely that these mechanisms will be perfect in equilibrium and so we assume that  $z_i$  is not empty. Finally, let respective maps to value estimates be the functions  $g$  and  $h$ :

$$(1) \quad \hat{V}_p^i = g(x_i);$$

$$(2) \quad \hat{V}_f^i = h(x_i, z_i),$$

where  $\hat{V}_p^i$  and  $\hat{V}_f^i$  are random variables that correspond to the public and private estimates of the value of lease  $i$ . The simple

form taken by (1) and (2) belies the potential complexity of the process under consideration. For example, the vector  $x_i$  contains objective data on production and information each firm reveals about the characteristics of its leases. Accordingly, misrepresentations concerning lease characteristics are also included in the  $x_i$ , though they may be completely discounted and not affect  $\hat{V}_p^i$ . Similarly, if firms truthfully reveal a characteristic, but it is not believed because it cannot be verified, that information also will not affect  $\hat{V}_p^i$ , but it will affect  $\hat{V}_f^i$ . On the other hand, if firms lie regarding lease characteristic  $j$ , and the lies are believed, then  $z_j^i$  will be the difference between the true value of the characteristic and the value  $x_j^i$  believed to be true by other parties. In this case the misrepresentations clearly influence  $\hat{V}_p^i$ . A related case is if firms truthfully reveal only favorable information, then  $z_i$  consists of the unfavorable information held back;  $x_i$  will not include unfavorable data available only from the firm, if the firm cannot be induced to reveal it. There are also other possibilities. Hence, while (1) and (2) take a simple form they admit the full range of value estimation problems due to asymmetric information. It is convenient to assume, for now, that both information sources are unbiased:

$$(3) \quad E(\hat{V}_p^i) = E(\hat{V}_f^i) = V_i,$$

where  $V_i$  is true lease value and  $E(\cdot)$  is the mathematical expected value. In principal, of course, either or both estimation procedures may be biased as discussed above. The simplifying assumption in (3) is that the estimation rule, (1), can correct for these potential biases. This assumption is made solely to illustrate that contracting conflicts can occur even when actors have rational expectations with respect to lease values. As it turns out, this assumption can be relaxed easily, but this would only complicate the exposition while adding little additional insight. Conflict over estimated lease values and unit shares is the heart of the contracting problem analyzed below, and it is not critical whether the conflict is due to inherent uncertainty about values or because of unrecognized biases in the estimation of lease values.

<sup>5</sup> The exact amount of information that can be communicated is a complex theoretical problem. Since firms have an incentive to selectively reveal only favorable information about the value of their leases, there will be some degree of imperfection in communication. Moreover, as described in the text, lease values estimates are influenced by firm-specific management policies, the details of which are difficult to reliably convey to other firms.

Finally we assume that unit shares will be based upon public estimates of lease values:

$$(4) \quad S_i = \hat{V}_p^i / V^*.$$

where  $V^* = \sum \hat{V}_p^i$ . This assumption is not central to the analysis and is used only to close the model.<sup>6</sup> Any alternative assumption regarding share assignment will yield the same qualitative results for the analysis so long as public value estimates influence final share assignments.<sup>7</sup> We will refer to (4) as the allocation rule. Given that the allocation rule will be the one offered the firm in unit negotiation, we can now characterize firms' decisions to join the unit.

The firm's objective in considering whether to join is to maximize the expected present value of its leases with respect to the date it joins the unit,  $t_u$ , and it may choose never to join. Formally, we have

$$(5) \quad \text{Max}_{t_u} E_f(PV_i) \\ = E_f \left\{ \int_{t_0}^{t_u} \pi_n(t) s_n^i(t) e^{-rt} dt \right. \\ \left. + \int_{t_u}^T \pi_u(t_u, t) s_u^i(t_u) e^{-rt} dt \right\},$$

where  $PV_i$  = the present value of lease  $i$ ,  $t_u$  = the time the lease is put in the unit,  $\pi_n$  = the stream of nonunitized net revenues from the field,  $s_n^i$  = lease  $i$ 's share of non-unitized net revenues,  $\pi_u$  = the stream of unitized net revenues from the field,  $s_u^i$  =

lease  $i$ 's share of unitized production,  $T$  = field life,  $r$  = the common discount rate for all firms, and  $E_f(\cdot)$  represents the firm's expectations. Differentiating (5) with respect to  $t_u$ , we derive the necessary first-order condition for the firm to join the unit at a particular time:

$$(6) \quad E_f \left\{ \left[ \pi_n(t) s_n^i(t) - \pi_u(t) s_u^i(t_u) \right] e^{-rt} \right. \\ \left. + \int_{t_u}^T \left[ \frac{\partial \pi_u(t_u, t)}{\partial t_u} s_u^i(t_u) \right. \right. \\ \left. \left. + \pi_u(t_u, t) \frac{\partial s_u^i}{\partial t_u} \right] e^{-rt} dt \right\} = 0.$$

Rewriting (6) we have

$$(7) \quad E_f(\alpha_i + \eta l_i) = E_f(\beta_i - s_i \lambda_i),$$

where  $\alpha_i = \pi_n(t) s_n^i(t) e^{-rt}$ ,

$$\beta_i = \pi_u(t) s_u^i(t_u) e^{-rt}, \quad l_i = \partial s_u^i / \partial t_u,$$

$$\eta = \int_{t_u}^T \pi_u(t_u, t) e^{-rt} dt, \quad s_i = s_u^i(t_u),$$

$$\lambda_i = \int_{t_u}^T \frac{\partial \pi_u(t_u, t)}{\partial t_u} e^{-rt} dt < 0.$$

The separation of  $l_i$  from the integral is appropriate since shares are fixed at the time of unitization.  $\alpha_i$  represents the firm's instantaneous nonunitized net revenues, and  $\eta l_i$  represents the gain in unitized share the firm expects to receive if there is delay in unit formation.  $\beta_i$  represents the firm's instantaneous share of unitized net revenues, and  $\lambda_i$  represents the firm's share of lowered field rents caused by delayed unitization.

Since  $\sum_i l_i = 0$  and  $\sum \alpha_i \leq \sum \beta_i - \sum l_i$ , there are aggregate incentives to unitize, but the division of the increased rents determines whether or not an individual firm will have incentive to join. Before considering reasons why, according to the model, certain firms will delay joining, let us rule out some other reasons for delay that are inconsistent with the model. Suppose the sharing rule is the number of wells, then the firm will not delay

<sup>6</sup>Given (4), the economic content of (3) follows without loss of generality. Since  $\sum S_i = 1$  by construction, the average share must be unbiased. Selective revelation or distortion of facts, then, is a zero sum game that can only increase the variance in the assigned shares about the true "fair" shares, i.e., those that would be assigned if  $V_i$  rather than  $\hat{V}_p^i$  were in the numerator of (4). Hence, if (3) fails, and the estimates are biased, rather than using the terminology "variance" below, "Mean Squared Error" should be substituted and similarly for higher moments, but otherwise the analysis is unchanged.

<sup>7</sup>For example, if unit shares in equilibrium were an increasing but strictly concave function of  $\hat{V}_p^i$ , because small firms can force concessions from large firms, the analysis below still applies.

unit formation merely to drill more wells absent disagreement over the number of wells it *can* drill. If all parties agree on the number of wells that can be drilled on a lease and if there is no desire to change the way in which wells enter the allocation rule, then the potential number of wells will enter  $\hat{V}_f^i$  and  $\hat{V}_p^i$  in precisely the same way, and there will be no need to delay unitization until the wells are actually drilled. Similar reasoning applies to any other strategic advantage of one party; so long as other parties recognize the valuation consequences of the advantage, it becomes immediately incorporated into the allocation rule, and delay is unnecessary.

There are other reasons, however, why the left-hand side of (7) may be greater than the right, in which case the model predicts that the firm will prefer to delay unit formation. A firm may wish to delay unit formation if its private value estimates exceed public value estimates,  $\hat{V}_f^i > \hat{V}_p^i$ , and if the firm expects the public estimate to be revised upward as more information on lease value becomes available. Differences between public and private information can occur for two reasons: (i) the functions  $g$  and  $h$  differ so that the mappings from raw data to value estimates are not the same; and (ii) there is private information,  $z_i$ , that causes the firm to believe public estimates are inaccurate. If the firm believes that the function  $g$ , used to map known data to value estimates, seriously underestimates its leases' future productive potential, then the firm will have incentive to delay unit formation by withholding its leases in the expectation that subsequent production data will cause an upward revision in assigned shares. Further, the firm's engineers and geologists have access to company records, the accuracy of which cannot be verified by outsiders. If these records cause them to believe that future lease output (value) has been underestimated ( $\partial h / \partial z > 0$ ), then there will also be incentive to delay unit formation to await future production information that will be expected to increase  $\hat{V}_p^i$ .

These information problems aside, the firm may also decide to delay joining, if holding out will alter the allocation rule, (4), through concessions from the other parties since de-

lay causes aggregate losses. Theoretically, then, firms may hold out in an attempt to delay unit formation either due to differences in value estimates, where  $\hat{V}_f^i > \hat{V}_p^i$ , or because they want to alter the allocation rule. If the firm expects the incremental gains in assigned share from new information and strategic bargaining to offset the firm's share of losses in field value due to delay, the firm will not join. Formally, the firm will not join if

$$(8) \quad E_f(l_i) > (\beta_i - \alpha_i - s_i \lambda_i) / \eta,$$

where  $E_f(l_i)$  is the firm's expectation of the change in its unit share due to delay.

The special case where instantaneous profits are the same under open and unitized production,  $\beta_i = \alpha_i$ , yields intuitive insight into the problem. In this case, (8) implies

$$(9) \quad \frac{E_f(l_i)}{s_i} \equiv \frac{\partial s_u^i}{\partial t_u} \bigg/ s_i > \frac{-\lambda_i}{\eta} \\ \equiv - \int_{t_u}^T \frac{\partial \pi_u(t_u, t)}{\partial t_u} e^{-rt} dt \\ \bigg/ \int_{t_u}^T \pi_u(t_u, t) e^{-rt} dt.$$

In words, if the firm's expected percentage gain in share exceeds the percentage loss due to delayed unitization, the firm will not join.<sup>8</sup> There exists an expected share revision,  $l_i^*$ , which just creates an equality in (9) and is the dividing line between firms that want to join and those that want to delay. Because the right-hand side of (9) is strictly positive, the threshold value  $l_i^*$  must also be strictly positive. The probability that the firm will wish to delay, then, is the probability that  $E_f(l_i) > l_i^*$ . Because both estimates of value are unbiased,  $E_f(l_i)$  will have zero mean, and the probability that it exceeds  $l_i^*$  will

<sup>8</sup>Strategic bargaining aside, a necessary condition for (9) to hold is that the firm's private information leads to higher expected lease value than that based on public information.



depend critically upon the uncertainty of the public value estimates. As more lease information becomes known,  $\tilde{V}_p^i$  collapses around  $V_i$ , and the returns to waiting for more favorable information vanish. Hence, the probability that  $E_f(l_i) > l_i^*$  decreases monotonically with the reliability of public estimates of lease value.

Assume that  $E_f(l_i)$  is normally distributed with zero mean and variance  $\sigma$ , and define  $p(J)$  as the probability that lease  $i$  will join at time  $t$ ; then

$$(10) \quad p(J) = p\{E_f(l_i) \geq (\beta_i - \alpha_i - s_i \lambda_i) / \eta\} \\ = \Phi(\psi / \sigma),$$

where  $\Phi$  is the normal distribution function and  $\Psi \equiv (\beta_i - \alpha_i - s_i \lambda_i) / \eta$ . As the variance  $\sigma$  declines, the probability of joinder increases.

Figure 1 illustrates the impact of more precise public value estimates on the probability that a lease will join. The figure shows the density function for the firm's expectation of the change in share,  $E_f(l_i)$ , and the mean is equal to zero since both estimates of value are unbiased. As the variance of public estimates of lease value declines, the probability to the right of the threshold change in share,  $l_i^*$ , declines and the probability of joinder rises.<sup>9,10</sup>

This leads to an important implication that leases with more uncertainty regarding their value will be less likely to join units early in negotiations. We show below that these leases tend to be those with the longest productive lives and the greatest estimated values. Over time, greater information on lease values becomes available, uncertainty

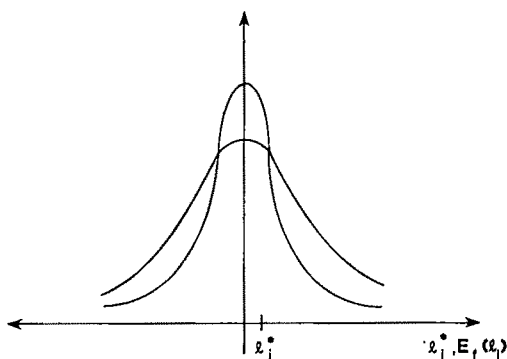


FIGURE 1

declines, and public and private estimates of lease values collapse around the true value. Accordingly, the model implies that unitization contracts are more likely to be completed late in the primary productive life of the reservoir.

Another factor affecting the probability of delaying unit information is the location of the threshold  $l_i^*$  itself. A major influence on this threshold  $l_i^*$  is the size of the firm's holdings on the field (see (9)). As size increases, the firm will bear a larger share of the cost of delayed unitization. Hence, for a large firm to postpone joining, the expected share increase from delay must be larger, moving  $l_i^*$  to the right. This implies that very large firms on a field are more likely to join early.

Finally, a firm may delay joining for strategic bargaining reasons as we have noted above. In such cases the firm attempts to extract share concessions as a means of gaining ratification of unitization. In such circumstances the firm expects a positive  $E_f(l_i)$ , *ceteris paribus*.

There are other potentially important contracting problems. First, there will be some opportunity for the unit operator to take advantage of his position due to imperfect monitoring by other firms. Opportunism provides additional incentive for the operator to put his leases in the unit, but is a source of concern to other firms. A second problem we have referenced above is the case where a firm not only delays, but decides not to join the unit when it is finally formed. The firm's

<sup>9</sup>This generalizes beyond the normal density to any mean-preserving spread for any density so long as probability weight is being taken from  $[0, l_i^*]$  and placed to the right of  $l_i^*$ .

<sup>10</sup>Other factors affecting  $E_f(l_i)$  are the rate of change of the public information stock and strategic bargaining. The arguments in the text apply, *ceteris paribus*, for a given rate of change of these variables. To ease exposition the example chosen is where there is zero expected share change due to strategic bargaining.

decision rule under these circumstances is

$$(11) \quad \int_{t_u}^T \pi_n^i(t) e^{-rt} dt \\ > \int_{t_u}^T \pi_u(t) s_u^i(t_u) e^{-rt} dt.$$

In this case an all-or-nothing offer is being made by the unit to the firm to, effectively, play a positive sum game. The offer made must reflect what the firm can achieve independently or it will not play. Further, if the offer exceeds by any amount what the firm can achieve independently then, in general, it will accept.<sup>11</sup> We show below that certain firms systematically refuse these offers. Empirically, they are not offered greater shares to join after the unit is formed. In this case there must be disagreement regarding what the firm can independently achieve. Central to the firm's viability outside the unit is whether or not it has a large enough segment of the field to operate effectively as a separate entity. This is particularly important for secondary oil recovery. With secondary recovery, injection of natural gas or water increases pressure and would drive oil to adjacent areas unless the firm drilled injection wells along the perimeter of its leases. Hence there are substantial effective economies of scale for these secondary recovery projects. For small acreage, the added costs of injection wells would likely swamp any gains from separate secondary recovery. Therefore, a firm with large contiguous acreage will be better situated to remain independent of the unit if his private information suggests that his leases have higher values than public information suggests.

One task of the empirical analysis in Section III is to separate, where possible, the contracting problems due to information issues from those arising from a simple hold-

out strategy. The empirical observation that some firms choose to remain permanently out of a unit allows us to comment on hold-out strategies. Hold-out behavior is unlikely to be a major factor in the firm's decision not to join when a unit is formed, since firms who refuse to join are not later offered more favorable shares. Hence, holding out to increase one's share is not the motivation. Further, holding out to free ride on secondary recovery projects is also not empirically important. Firms that hold out often form their own subunits with neighboring firms, and cooperatively drill water injection wells along boundaries common with the unit to prevent oil migration between secondary recovery projects. This suggests that a firm's decision not to join is more complex than strategic bargaining to raise rental shares.

The theoretical discussion leads to a number of hypotheses to be tested in the empirical analysis in Section III:

1: As fields age, public and private information sources converge, making agreement on an allocation rule more likely.

2: Firms with large, scattered holdings on a field will be more willing to join and will also be more flexible in voting on allocation rules.

3: Similarly, firms with large blocks of contiguous acreage are more likely to withdraw and form separate units.

4: Leases where there is greater uncertainty of public estimates of lease value will be less likely to join a unit. It will be shown that high-output leases fit into this category.

5: Negotiation of allocation rules for assigning unitized rental shares will be constrained by disputes over estimates of structural characteristics.

6: Allocation rules must assign once-and-for-all shares, with no contingent updates.

7: Finally, the unit operator will be more likely to place a given lease in the unit, *ceteris paribus*.

The empirical analysis addresses the evidence concerning each of the (alternative) hypotheses. The discussion will generally be organized by hypothesis number.

<sup>11</sup> This decision is made recognizing that the firm will not be later offered a more favorable share. Such an offer would require a general renegotiation of shares, which is not possible. In this case, an all-or-nothing offer is made by the unit to the firm.

### III. Contracting on Seven Fields

The empirical bases for the study are unitization negotiations on seven Texas and New Mexico oil fields. Descriptive statistics for these fields and their unit negotiations are provided in Appendix A. Our examination of unitization contracting begins with a quantitative analysis of firms' decisions to place particular leases in the unit on the three oil fields for which we have sufficient data. The fields are North Cowden, Goldsmith/Landreth, and Prentice Northeast.<sup>12</sup> The theory presented above establishes the following partial derivatives for the probability that a given lease will be placed in a unit.<sup>13</sup>

$$(12) \quad J_{i,k} = f(A_i, C_{i,k}, Q_{i,k}, w_i, \varepsilon_{i,k})$$

$$f_1 > 0, f_2 < 0, f_3 < 0, f_4 > 0,$$

<sup>12</sup> The North Cowden field was discovered in 1930 and negotiations for the unit began in 1958 as the field neared depletion of primary reserves. An Engineering Committee was formed to collect parameter data from operators for assigning shares and to estimate gains from secondary recovery. In 1960 the committee estimated that recovery would increase by 100 million barrels, a gain of \$285 million in 1960 prices (Minutes, July 28, 1960, North Cowden Unit File, Company Records). Nevertheless, 19 of the 31 operators eventually withdrew part or all of their leased acreage, and a smaller unit was not formed until 1966, eight years after negotiations began. Contracting for the Goldsmith/Landreth unit began in 1961, but conflict over unit boundaries delayed work by the Engineering Committee until 1963. Four of the 10 operators bargaining for the unit withdrew leases in disagreement over allocation rules. Additionally, after a formula was finally agreed upon one firm withdrew eight more leases due to a dispute over secondary recovery plans. The final unit was not formed until October 1965. The Prentice field was discovered in 1951, and unit negotiations began in early 1954. Despite predictions that early unitization would substantially increase recovery, negotiations faltered, and were abandoned between 1956 and 1959. At that time the largest operator in the field attempted to reopen negotiations. By 1963 it was clear that field-wide unitization was not possible, and in late 1963 three units were formed, nearly ten years after negotiations were opened; the Northeast and Southwest units were operated by one firm and the central unit was operated by another.

<sup>13</sup> Strictly speaking, the hypothesis that  $f_3 < 0$  requires evidence (to be presented below) that increases in

where  $J_{i,k} = 1$  if lease  $k$  of firm  $i$  is placed in the unit, zero otherwise;  $A_i$  is firm  $i$ 's total acreage in the field;  $C_{i,k}$  is firm  $i$ 's acreage contiguous to lease  $k$ ;  $Q_{i,k}$  is the current output per acre on lease  $k$ ;  $w_i = 1$  if firm  $i$  is the unit operator, zero otherwise;  $\varepsilon_{i,k}$  is the error term; and where  $f_1 = \partial f / \partial A_i$ , and similarly for  $f_2$ ,  $f_3$ , and  $f_4$ . This equation can be used to test hypotheses 2, 3, 4, and 7. The assumption above that  $E_f(l_i)$  is normally distributed, and the binomial nature of the dependent variable requires a Probit estimation procedure for the joinder equation. Empirical results are presented in Table 2, part A. All coefficients are of the right sign and most are significant. Thus, the results generally provide strong support for the four hypotheses.

We have argued that firms with very productive leases will not place those leases into the unit. There is generally greater uncertainty about their value and disagreement over their share in the unit. The regression coefficients in Table 2, part A, are consistent with this notion. All of the output per acre coefficients are significant at the 10 percent level, and two are significant at the 1 percent level. Moreover, in Appendix B we show that there is greater uncertainty regarding predicted output levels for high output leases. We do this by regressing current output on all known structural characteristics and then examining the residuals. Those leases where uncertainty is highest will have the largest residual variance after adjusting for all factors known to affect output. We show that those leases with the highest output have larger unexplained (residual) variance. According to the theory, this should increase the probability that such leases will be withheld from the unit. The results of the estimations in Table 2 strongly support this hypothesis.

Qualitative evidence also indicates that the pattern of withholding very productive leases is consistent on all seven fields, despite ag-

lease output lead to greater uncertainty in value estimates.

TABLE 2<sup>a</sup>

	Intercept	Unit Operator	Lease Output per Acre	Firm Acres Contiguous to Lease	Total Firm Acres
A. Probit Estimations of Joinder Decisions <sup>b</sup>					
North Cowden	-.51 (.34)	11.7 (34.2)	-.0051 (.0032)	-.00096 (.00025)	.00050 (.00024)
$\chi^2 = 54.57$					
Goldsmith/Landreth	2.70 (1.01)	13.6 (13.5)	-.0118 (.0040)	-.0061 (.0014)	.00061 (.00063)
$\chi^2 = 39.49$					
Prentice Northeast	.736 (.569)	4.94 (74.6)	-.0062 (.0028)	-.0130 (.0044)	.0120 (.0042)
$\chi^2 = 33.18$					
B. Voting Patterns on Empire Abo Unit <sup>c</sup>					
	9.67 (2.74)				
$F(1, 11) = 7.16$					.0066 (.0025)

<sup>a</sup>Standard errors are shown in parentheses.

<sup>b</sup>Dependent variable = 1 if lease is placed in unit; 0 otherwise.

<sup>c</sup>Dependent variable: the number of yes votes by firm in unit balloting (58 ballots).

gregate gains from a field-wide unit. On the Prentice field, a group of the most productive leases were located in the center of the field. After nine years of negotiations, efforts to form a single unit were abandoned because "a common formula could not be negotiated" (Minutes, Operators' Meeting, February, 1963, Prentice N.E. Unit File, Company Records). As a result three units were formed with separate secondary recovery projects. On the North Cowden field, firms with very productive leases voiced opposition to the proposed unit, with one firm asserting "none of the proposed parameters give justice to those leases because of their abnormal producing capabilities" (Letters, January 20, 1959; March 30, 1961, North Cowden Unit File, Company Records). Eight of the firms withdrawing acreage had ten of the most productive leases on the field. Output on those leases for the first six months of 1960 averaged 133 barrels per acre, while the average for all other leases was 79 barrels per acre (Engineering Report, December 1, 1960, North Cowden Unit File, Company Records). One unit proponent reported: "It is extremely difficult to arrive at a single factor which can be said to represent an equitable minimum, since the field is currently under active development and current production

relationships are changing with each month's data" (Letter, February 5, 1960, North Cowden Unit File, Company Records). Similarly, on Goldsmith/Landreth, three firms with unusually productive leases requested that acreage be deleted from the allocation rule. The three leases involved had average output per acre of 233 barrels for the period June 1962 to June 1963; only one other lease on the field had production of 200 barrels per acre, and average output for all other leases was 80 barrels, about one-third that of the withdrawn leases (1963 Engineers' Report, Table 4, Goldsmith/Landreth Unit File, Company Records). On the seven fields, owners of very productive acreage typically stressed that none of the formulas under consideration adequately protected their equity. Despite long negotiations (see Appendix A) and repeated formula adjustments (58 different votes on Empire Abo alone), consensus could not be reached on lease values and unit shares to attract the most productive leases into the unit. Thus, highly productive leases were systematically segregated from less productive leases and, where possible, separate subunits were formed.

When separate subunits were formed, independent secondary recovery projects were undertaken that were typically less effective

than field-wide projects. This indicates that firms could not agree on relative shares, since they were willing to bear higher costs. Moreover, there was no prospect of later renegotiation, and neither party was able to free ride on the other's secondary recovery efforts. Hence, parties disagreed about the relative values of leases, leading to a breakdown in negotiations. These observations suggest bargaining problems that are more complex than a simple hold-out problem.

Hypothesis 1 states that as field-wide primary depletion nears, consensus on unitization is more likely. The qualitative evidence from all seven fields supports this notion. Negotiations on North Cowden took eight years, in part, because some of the field was newly developed, while other parts were sharply declining. Much of the conflict centered on differences between new and older sections of the field. In withdrawing its lease from the unit, one firm notified the unit organizers that "the various parts of the field were simply so diverse that not one formula could satisfy everyone. We wish you every success in forming a unit in the center of the field where everything is more uniform" (Letter, September 4, 1963, North Cowden Unit File, Company Records). Similarly, agreement could not be achieved early in 1967 on Empire Abo when most leases could produce at regulated maximum production. Agreement could not be reached until 1971 as primary production declined.

The regression results in Table 2, part A, also support the notion that firms with large holdings on a given field will support unitization (hypothesis 2). This effect is measured by the total acreage variable, and the coefficients indicate that such firms were more willing to place their leases into the unit. These results are also backed by qualitative evidence. Firms with large holdings frequently made concessions to complete unit agreements. For example, on North Cowden, the largest firm estimated its 1959 output share at 37.5 percent of field production, but was willing to accept a lower minimum unit share of 36 percent (Letter, January 1, 1960, North Cowden Unit File, Company Records). On Goldsmith/Landreth, the concern of the largest firm regarding an increase in

unit costs due to the withdrawal of some leases is reflected as follows: "Although our reserves in the area from which all eleven tracks are eliminated are indicated to be greater..., our costs will undoubtedly be greater due to the requirement of additional injection wells..." (Letter, October 12, 1964, Goldsmith/Landreth Unit File, Company Records).

It was also predicted that firms with large acreage will support more allocation rules than will firms with small holdings. Data for 58 recorded ballots for 13 firms in Empire Abo negotiations allow for quantitative tests. In Table 2, part B, OLS regression results are reported for voting on Empire Abo:

$$(13) \quad Y_i = f(A_i; \eta_i); \quad f_1 > 0,$$

where  $Y_i$  = the number of yes votes by firm  $i$  in the balloting;  $A_i$  is firm  $i$ 's total acreage on the Empire Abo field; and  $\eta_i$  is the error term with zero mean, and it is assumed that  $A_i$  and  $\eta_i$  are independent. The test, then, examines the link between the total number of yes votes in repeated balloting and acreage. The coefficient for acreage is positive and highly significant, and these results are repeated elsewhere. On the Goldsmith San Andres field, the larger operators offered to give more than proportionately in share negotiations to speed the unit. In internal negotiating documents the firms recognized that they had 72.23 percent of current output, but they agreed to an aggregate share of 71.09 percent of remaining primary production under the unit and 67.80 percent of secondary recovery (Letter, October 27, 1961, Goldsmith San Andres Unit File, Company Records).

Firms with limited holdings, on the other hand, were more selective, supporting only specific allocation formulas that emphasized characteristics favoring their leases. For example, on the Slaughter Estate unit, one firm with 4 percent of acreage withdrew because the current output parameter used in the allocation formula supposedly undervalued its two leases. The firm had 2.81 percent of current production, but asserted that it had 3.53 percent of remaining oil reserves based on Engineering Committee estimates.

Two other firms with 4 and 2 percent of acreage, respectively, also threatened to withdraw for similar reasons (Letters, December 26, 1962, June 6, 1963, Slaughter Estate Unit File, Company Records). On Goldsmith San Andres none of the firms consistently voting no on allocation formulas had over 9 percent of field productive acreage. One small firm with only .3 percent of acreage voted no on all formulas offered. These small firms repeatedly called for adjustments in the weights placed on specific parameters to reflect their individual advantages: one firm with 5 percent of acreage and 1 percent of cumulative output called for less weight on the latter; another, with 4 percent of acreage and 2.8 percent of current output wanted current production removed or discounted. On the other hand, the three largest firms with 24, 16, and 15 percent of acreage, respectively, voted yes on all of the allocation rules submitted for consideration (Minutes, January 10, 12, 1962; February 7, 1962, Goldsmith San Andres Unit File, Company Records). In general, in contrast to the case where firms refused the final offer to join, evidence on voting behavior can reflect both information issues and hold-out strategies to increase rental shares.

The regression results in Table 2, part A, also support the notion that firms with large blocks of *contiguous* acreage will permanently withdraw and form their own units (see hypothesis 3, above). The coefficient for contiguous acreage is negative in all three fields and significant in two. On North Cowden, firms with large tracts of contiguous leases withdrew to form separate units. They included the second, fourth, and sixth largest firms by acreage on the field, and the withdrawn leases represented 97, 93, and 71 percent of the total acreage of these firms (Table 2, Engineers' Report, December 1, 1960, North Cowden Unit File, Company Records). On Western RKM, the second largest firm with 26 percent of acreage, all concentrated in the eastern part of the field, withdrew and formed its own secondary recovery unit.<sup>14</sup> The firm argued that the

parameters considered for share assignment "considerably underestimated" its lease values. The firm, along with other adjacent leases, formed its own secondary recovery unit (Letter, May 20, 1964, Western RKM Unit File, Company Records). On Prentice, the primary advocate of separate units had acreage of sufficient size concentrated in the center of the field for secondary recovery to be possible. In contrast, the unusually productive leases that were isolated in the northeast portion of the field finally joined the northeast unit; in part, because they were not large enough to be independently viable. As noted above, these decisions to remain outside a unit must be largely due to information issues and not a bargaining strategy.

The final coefficients in Table 2, part A, to be discussed are those for the unit operator variable. Hypothesis 7 is that the firm chosen to be unit operator would be more likely to place its leases into the unit. The regression results do not support this view; holding lease productivity, firm size, and contiguous acreage constant, the unit operator is not significantly more likely to place a given lease in the unit on all three fields.

There are two hypotheses not directly addressed by the regression estimates in Table 2. Hypothesis 5 is that the share parameters are limited because they must be based upon public information, and hypothesis 6 states that allocation rules must assign shares for all future periods. Qualitative evidence allows us to examine these hypotheses. The selection and interpretation of formula parameters were the central source of dispute on all seven fields. In general, agreement can only be reached for formulas using parameters that could be measured and interpreted without controversy. This sharply limits contractual flexibility because of the small number of objectively measurable variables and the highly tenuous nature of even modest extrapolations as discussed in Section I. Even simple static structural char-

<sup>14</sup>In March 1962, a 26,400 acre RKM unit was planned on the Slaughter field, but was soon dropped in

favor of a smaller 9,911 acre unit. Negotiations continued through 1966. By that time 50 percent of the leases, covering 4,993 acres, had been withdrawn. Most were on the eastern side of the proposed unit, leaving only 4,918 acres in the final agreement.

acteristics, such as the thickness of the reservoir rock and pore space available for holding hydrocarbons, were sources of significant dispute. For example, early in unit negotiations the Engineering Committee on North Cowden reported that data were too sketchy to calculate "a fair and equitable" gross or net pay under each lease (the gross thickness of the reservoir or the thickness net of any nonproductive zones). The Committee only had well cores for 28 of the 733 wells on the field, and it stressed the "meager data and poor quality of available records" (Memo, April 7, 1959, North Cowden Unit File, Company Records). As a result, during the eight years of negotiations for the North Cowden unit, nearly all of the numerous parameter formulas considered were simple convex combinations of current and cumulative output, which were available and reliable for all leases. Attempts to incorporate more sophisticated parameters met with objection due to their subjective nature given the lack of available data: "a disadvantage [of gross pay] is that we are basing a parameter of unitization on the skill or lack of skill in the persons observing the samples. Therefore, there is considerable question as to the consistency of the picks [estimates] between wells, and it would likely be difficult to reach agreement between operators on such data..." (Letter, June 16, 1959, North Cowden Unit File, Company Records). Moreover, remaining reserves could not be estimated in ways acceptable to all operators. The unit operator had access to reservoir data and could have estimated the parameter, but its estimates would have been controversial. Instead, it chose to release the data to the Engineering Committee for parameter calculation. Even so, one small firm hired an outside consultant to calculate its net reserves, and got values double those calculated by the Engineering Committee (Letters, January 9, 1962; March 8, 1963, North Cowden Unit File, Company Records).

Disputes over the measurement of static reservoir characteristics occurred as well on other fields. On Prentice Northeast, the owner of the most productive acreage believed that its reservoir pore space had been badly underestimated. Reconciliation took several

months and finally resulted in an *ad hoc* upward adjustment of 70 percent, and led to a large increase in the firm's unit share. A major information problem was that observations were limited to wells, and the extrapolation method used between wells dramatically altered parameter calculations. For example, a linear versus a log-linear extrapolation of reservoir depth between wells created significant differences in estimated reservoir volume under particular leases. When the reservoir was highly variable or when observations for a large percentage of wells were unavailable, as on Cowden, the issues were extremely difficult to resolve. On Prentice Northeast, the final formula accepted included one observable and measurable variable, current production. Estimates of primary recovery for the leases were made using a variety of techniques, but "Relatively poor agreement between various methods were obtained in many instances. Primary reason for inconsistencies [was] due to a severe lack of control in much of basic data together with inherent uncertainties involved in these type calculations." Core analyses done by differing consulting firms gave dramatically different results (Engineering Committee Minutes; Letter, June 29, 1963, Prentice N.E. Unit File, Company Records). These data problems persisted in 1963, even though Prentice Northeast was fully developed and two-thirds depleted. The inability to precisely estimate reserves at the field level is not a serious obstacle to contracting. Disputes about reserves and future productivity, however, are repeated at the lease level, where they break down share negotiations.

Finally, as predicted in hypothesis 6, the accepted formulas for each unit assigned *permanent* shares. Contingent updates are not feasible because unitization fundamentally changes the pattern of reservoir production. Some wells are plugged, others are converted to injection, and new wells are drilled, completely altering the pattern of migration. Hence, after unitization it is impossible to infer the oil in place or the oil that could have been produced from a given lease. In no case have we encountered a unit where contingent updates were allowed. Evidence of the once-and-for-all nature of the con-

tracts is that the allocation formulas adopted were often multiphase, but established a set pattern of share adjustment, based on data known at the time of agreement. Because unit shares were fixed, those firms most concerned about biases in their fair share calculation due to incomplete information would be reluctant to join, based on existing evidence. We have identified those firms as owners of very productive leases, and their tendency to withdraw from unit contracting has been documented above. If contingent updates were possible, negotiations could allow for temporary shares to entice early agreement with corresponding aggregate gains, and share adjustments could be negotiated as more information became available.

Before turning to the conclusions, it is worth briefly reviewing the hold-out issue. Previous studies of unitization (see, for example, Stephen McDonald) have casually concluded that the primary obstacle to unitization is simple strategic bargaining to increase unit shares, rather than real differences in opinion regarding relative lease values. Clearly an element of both may be present in a given situation, and in many contexts it is difficult to distinguish between "honest" differences and pretended differences to gain an advantage. In the present context, however, there is evidence presented above that shows information problems play a critical role in contractual failure for unitization. The first of these is that the firms' joinder decisions involved a decision to permanently withdraw from the unit. The purpose of a strategic hold out is to gain a more favorable offer later; yet, for all of the decisions represented by the regressions in Table 2, part A, the decisions were once and for all. For these cases, then, strategic bargaining by either party was not the point. Second, a pure strategic hold-out hypothesis also would predict that high- and low-output leases are equally likely to delay the unit or refuse to join. The systematic differences in behavior evident between these groups is inconsistent with a simple hold-out hypothesis, and yet follows naturally from the imperfect information argument. Furthermore, breakdown in unit negotiations often led to the formation of multiple subunits on fields such

as Prentice. Here, the problems of reconciling differences in the value of highly heterogeneous leases were repeatedly cited. The operators, in explaining why they would not (and later *did not*) join a unit, claimed that the heterogeneities could not be reconciled. These same operators, then, contemporaneously formed partial units with other operators having similar, neighboring leases. A hold-out strategy does not explain why these operators would withdraw from one unit and form another.

Third, we have a very interesting set of observations from unitization on federal lands in Wyoming. The federal government actively encourages unitization through policies that facilitate agreement (see our earlier paper).<sup>15</sup> On federal lands, agreements can be reached prior to the drilling of any wells—a period when *private* information about lease characteristics such as well performance will be limited and known lease heterogeneities are minimal. Units formed at this time take less than six months to negotiate, and the percentage of production that is unitized is very high. Further, there is little dispute chronicled in the negotiation records. If, however, the unit is formed later after production has begun, and lease heterogeneities and asymmetric information have emerged, then negotiations take an average of seven years and are frequently acrimonious. Hence, the federal policies work well before asymmetric information and heterogeneities have emerged, but have little impact if agreement is not reached during that period. A simple hold-out strategy cannot easily explain why the commencement of production fundamentally changes the character, speed, and success of unit contracting.

None of this is to suggest that hold outs are an unimportant problem in unitization. Instead, the analysis of this paper shows that information problems play a critical role,

<sup>15</sup> General policy responses to the unitization problem are beyond the scope of this paper. Our forthcoming paper presents analyses of the impact of different unitization policies in Oklahoma, Texas, and federal lands on the speed and extent of unitization. We also offer reasons for sharp policy differences across these political jurisdictions.



and explain much observed behavior that is otherwise difficult to unravel.

#### IV. Concluding Remarks

Private contracting is a common solution to many problems in production and exchange. When there exists a core of exchange, contracts permit parties to move into the core, and general welfare is improved. Recent analyses of contracting by Goldberg, Williamson, and others, however, have introduced a point of caution—information asymmetries, opportunism, and small numbers bargaining problems can break down the contracting process in cases where there is a need to sequentially update contract terms to reflect changes in the economic environment. These findings limit the range of problems where contracting can be an effective solution in resource allocation.

Our results have more serious ramifications for the general applicability of contracts. The study has analyzed an important empirical setting where private contracting has not been successful. Despite large net gains from unitization, *ex ante* imperfect information and information asymmetries among the negotiating parties regarding lease values prevents consensus on unit shares. These problems exacerbate any hold-out strategies that would otherwise impede agreement. As a result, contracts are often either not completed or are very incomplete with only fragmented units. Even simple once-and-for-all contracts that need little subsequent adjustment systematically break down. While one can think of a variety of hypothetical mechanisms for eliciting the information needed for agreement in an incentive compatible form, they are not observed.

We believe that the informational imperfections and asymmetries that lead to contractual failure in unitization are repeated in many contexts. Examples include labor markets and the problems of evaluating heterogeneous workers and contracting for pollution control. Our analysis suggests that an assumption that private contracting will solve inefficiencies in these areas is unwarranted. Accordingly, close attention by economists

to contracting details in a variety of settings is required for a better understanding of economic processes and events.

#### APPENDIX

##### A. Contracting Summary

Table A1 outlines the general contracting conditions for the seven fields examined in the paper. The table reveals the long bargaining time before agreements were reached. Moreover, the table shows that in every case but Empire Abo a subunit was formed, rather than a complete field-wide unit. Empire Abo is on federal land in New Mexico where compulsory unitization rules apply to force nonjoinders into the unit.<sup>16</sup>

##### B. Lease Output and Uncertainty of Lease Value

Here we demonstrate that highly productive leases have more uncertainty regarding their value than do less productive leases. Uncertainty is measured with respect to the publicly available data regarding lease characteristics. Greater uncertainty in public estimates of lease value implies, *ceteris paribus*, a greater probability of relatively large changes in expected share ( $E_f(l_i)$ ). In other words, the probability to the right of  $l_i^*$  in Figure 1 is larger. This makes it less likely that the lease will be placed in the unit. We test this proposition regarding high output leases in Section III.

In order to establish greater uncertainty in public estimates of lease value, we examine data concerning current output and all known basic reservoir characteristics. The argument is that inability to accurately predict current output using all known lease characteristics is equivalent to being unable to predict future output and, hence, current value. What we wish to show is that there is greater residual uncertainty regarding current output for highly productive leases than for less productive leases.

<sup>16</sup> Compulsory unitization rules do not exist in Texas; see our forthcoming paper.

TABLE A1

Field	Discovery Date	Date Unit Negotiations Began	Time until Unit Formed	Acreage Under Negotiation	Acreage in Final Unit	Number of Operators
North Cowden	1930	1958	8	30,870	17,503	31
Goldsmith/Landreth	—	1961	4	10,760	8,985 <sup>a</sup> 7,815	10
Prentice						
Northeast	1951	1954	9	8,500	6,828	—
Western RKM	—	1962	4	16,400	4,918	16
Slaughter Estate	—	1958	5	5,528	5,280	4
Empire Abo	1957	1965	6	11,323	11,323	15
Goldsmith San Andres	—	1959	4	7,199	6,103	27

Sources: Compiled from Unit Files, Company Records.

<sup>a</sup>Two reservoirs.

TABLE A2—HETEROSCEDASTICITY REGRESSION

	Independent Variables <sup>a</sup>					
	C	Pore Feet per Acre	Total Feet per Acre	Oil Originally in Place	Wells per Acre	Cumulative Output per Acre
Prentice						
(A)	-1.16 (-.16)	27.74 (.99)	-4.78 (-2.99)	4.43 (.91)	2445 (2.99)	.037 (4.65)
(B)	98.1 (.19)	2114.88 (1.48)	-144.1 (-1.76)	-293.4 (-1.18)	143700 (3.42)	.177 (.42)
Cowden						
(A)	46.1 (1.28)		-1.30 (-1.82)		.009 (6.19)	662 (.45)
(B)	-140 (-.09)		-2.06 (-1.36)		79625 (1.28)	(0.008) (0.138)
Goldsmith						
(A)	28.3 (2.89)				33.7 (.17)	.036 (5.75)
(B)	367.1 (.40)				824 (.04)	1.55 (2.63)

Note: Equation (A): Dependent variable = Output per acre; Equation (B): Dependent variable = Squared residual from Equation (A).

<sup>a</sup>t-values are shown in parentheses. Independent variables are different across fields solely because of differing data limitations in unit negotiation records. In all cases, the available independent variables used were those not calculated exclusively from current output.

To show this we perform a simple heteroscedasticity test. We first regress current output on *all* known lease characteristics, such as past output, number of wells, and reservoir thickness. We then examine the residuals of this regression. The null hypothesis is that the residuals are homoscedastic, while the alternative is that leases with higher output will have larger residuals than those with lower output. To test for this we estimate a second regression with the squared residuals

from the first regression on the left-hand side and the same set of variables on the right. The null hypothesis is rejected if all variables that are statistically significantly different from zero in both regressions also have the same sign in both regressions.

The regression results are reported in Table A2. For each field the null hypothesis of homoscedastic residuals is rejected at the 10 percent confidence level and for two of the three fields the null hypothesis is rejected at

the 5 percent confidence level.<sup>17</sup> Hence, there is greater residual uncertainty regarding prediction of current output from known lease characteristics for high-output leases than for low-output leases. Thus there are wider confidence bands on public estimates of output for these very productive leases. This increases the probability, *ceteris paribus*, that firms will have private information that causes them to believe that a substantial revision in share will be forthcoming, if the unit is delayed. In a Bayesian sense the strength of the prior, based upon public information, is less. Similarly, it increases the probability that they will not join a given unit because they believe their property is sufficiently undervalued that they can do better on their own. This proposition is tested in Section III.

<sup>17</sup> Under the null hypothesis, these tests are unbiased.

#### REFERENCES

- Bain, Joe, *The Economics of the Pacific Coast Petroleum Industry, Part III*, Berkeley: University of California Press, 1947.
- Coase, Ronald, "The Problem of Social Cost," *Journal of Law and Economics*, October 1960, 3, 1-44.
- Goldberg, Victor, "Regulation and Administered Contracts," *Bell Journal of Economics*, Autumn 1976, 7, 426-52.
- Libecap, Gary D. and Wiggins, Steven N., "The Influence of Private Contractual Failure on Regulation: The Case of Oil Field Unitization," *Journal of Political Economy*, 1985 forthcoming.
- McDonald, Stephen, *The Leasing of Federal Lands for Fossil Fuel Production*, Baltimore: Johns Hopkins University Press, 1979.
- Myers, Raymond M., *The Law of Pooling and Unitization: Voluntary and Compulsory*, 2d ed., Albany: Banks, 1967.
- Startz, Richard, "Prelude to Macroeconomics," Department of Finance, University of Pennsylvania, 1983.
- Williamson, Oliver, "Franchise Bidding for Natural Monopolies—In General and with Respect to CATV," *Bell Journal of Economics*, Spring 1976, 7, 73-104.
- U.S. Bureau of Mines, *Annual Report of the Director*, Washington, 1916.

# Anticipated Devaluations, Currency Flight, and Direct Trade Controls in a Monetary Economy

By ROBERT C. FEENSTRA\*

While a devaluation is the usual recommendation for a country facing a balance of payments deficit, this advice is often resisted by country governments. Instead, direct trade controls such as tariffs, quotas, and subsidies are frequently used. In this paper I propose an economic explanation for the reluctance to devalue and the use of trade controls, namely, the "currency flight" that can be expected to occur if the devaluation is anticipated. In an effort to avoid or offset the effects of this currency flight, countries may use direct controls. I shall analyze the properties of using import tariffs and export subsidies, with or without an anticipated devaluation, to affect the balance of payments and obtain the social optimum.<sup>1</sup>

Specifically, I shall identify three reasons for using direct trade controls when faced with a balance of payments deficit:

- 1) An equal import tariff and export subsidy will lower consumption and improve the balance of payments through a real balance effect, but will not cause speculative activity so long as goods are nondurable;

- 2) Prior to an anticipated devaluation we expect a depreciation of the black market exchange rate, and an import tariff is needed to offset the incentive to hoard foreign exchange and consume foreign goods;

- 3) When a social foreign exchange constraint is binding, a planner will shift consumption away from the import good, and a

long-run import tariff is needed to ensure that consumers follow this socially optimal plan.

The equivalence between a uniform tariff-cum-subsidy policy and a devaluation is well known (see Richard Caves and Ronald Jones, 1981, p. 355, and Gerald Meier, 1980, p. 170). It is also clear that with nondurable goods the tariff-cum-subsidy will not cause speculative activity, in contrast to the anticipated devaluation.<sup>2</sup> This is my first reason for using tariffs and subsidies. Second, in the absence of exchange controls, the currency flight caused by an anticipated devaluation will drain the central bank of reserves and create a black market in foreign exchange, and I shall solve for a depreciation of the black market exchange rate. The depreciation causes the consumer to substitute foreign for domestic money in asset holdings, and consume relatively more of the foreign good. This distorting effect on consumption can be offset by using an import tariff.

The third reason I identify for using trade controls is more specific to the model structure. In Section I, I describe a small country, intertemporal trade model where money is introduced by a continuous time version of "cash-in-advance" constraints, as originally proposed by Robert Clower (1967). It is assumed that the consumer uses domestic (foreign) money to purchase domestic (foreign) goods, and thus faces two cash-in-advance or transactions constraints. By endowing the consumer with large holdings of

\*Department of Economics, Columbia University, New York, NY 10027. I have benefited from the comments of seminar participants at Carleton, Columbia, Minnesota, Princeton, and Rochester universities. Special thanks go to Ronald Findlay for suggestions that have significantly improved this paper.

<sup>1</sup>Using the equivalence shown by Jagdish Bhagwati (1969), the import tariffs could be replaced by import quotas. In this sense, I do not maintain any distinction between price and quantity controls.

<sup>2</sup>This contrast between a tariff-cum-subsidy policy and anticipated devaluation can be further understood by recognizing, as in Bhagwati (1968, p. 63), that for full equivalence between the two policies the tariffs and subsidies should apply to the current and capital account. Since the direct trade controls I analyze do not apply to the capital account, and goods are nondurable, these controls do not cause speculative activity.

domestic money, the desired consumption level is high and a balance of payments deficit occurs. Various policies to obtain balance of payments equilibrium, such as equal tariffs and subsidies or a devaluation, are examined in Section II.

In Section III, I argue that the social planner faces *only* a transactions constraint in foreign exchange, since any amount of domestic money could in principle be created to avoid the domestic constraint. It follows that the policies analyzed earlier to secure balance of payments equilibrium are not sufficient to obtain the social optimum. In Section IV, I examine methods to obtain balance of payment equilibrium and decentralize the optimum. In particular, it is found that an import tariff greater than the export subsidy is needed when the social foreign exchange constraint is binding, giving the third reason for the use of trade controls. In the concluding section, I discuss the relation of the analysis to existing literature. The formal proofs of lemmas and propositions have been omitted for brevity, but are contained in a working paper, available on request.

### I. The Model

On the production side, let us assume a Ricardian economy, where every unit of labor produces one unit of the export good. The foreign exchange price of the exportable, importable, and domestic money are all normalized at unity under *laissez-faire*. Both goods are nondurable. The representative consumer faces the transactions constraints

$$(1) \quad M \geq \alpha C_1, \quad M^* \geq \alpha C_2,$$

where  $M$  and  $M^*$  are holdings of the domestic and foreign money, respectively;  $C_1$  and  $C_2$  are consumption of the exportable and importable; and  $\alpha$  is a parameter which can be interpreted as the length of time money must be held to finance consumption.<sup>3</sup> Discrete time versions of these

cash-in-advance constraints applied to international trade models have been used by Elhanan Helpman (1981a, b), Helpman and Assaf Razin (1979), Robert Lucas (1982), Torsten Persson (1984), Alan Stockman (1980), and Lars Svensson (1985); recent surveys of autarky and trade models with such constraints are provided by Meir Kohn (1984), and Maurice Obstfeld and Stockman (1984, Sec. 5).

I shall assume that consumers are permitted to instantaneously exchange any amount of domestic money for foreign money with the central bank, at the price of unity. Thus, I am assuming no exchange controls. This framework clearly exaggerates the availability of foreign exchange in many developing countries. It has the advantage, however, of leading to a particularly simple form of currency flight: in response to an anticipated devaluation, consumers will exchange domestic for foreign money with the central bank until  $M = \alpha C_1$ , and thereby drain some of the government's foreign exchange reserves. In the following section I shall consider the implications of the consumer's demand for foreign exchange exceeding available reserves.

I shall further assume that there are no domestic or foreign bonds, so the consumer's assets are given by  $A \equiv M + M^*$  (recalling that the exchange rate is unity). The absence of assets other than money is mainly used to simplify the model, but also exaggerates the limited extent of capital markets in many developing countries (see Ronald McKinnon, 1973). Assets  $A$  are a state variable of the system, and changes over time according to  $\dot{A} = L - C_1 - C_2$ . The variable  $L$  is the fixed quantity of labor supplied and equals income with the wage equal to the exportable price of unity, while  $C_1 + C_2$  is the value of expenditure.

Denoting the instantaneous utility function by  $U(C_1, C_2)$ , the representative con-

<sup>3</sup>Formally, I could write the transactions constraint for domestic money as  $M(t) \geq F(\alpha) \equiv \int_t^{t+\alpha} C_1(s) ds$ ,

and similarly for foreign money. A Taylor-series expansion gives  $F(\alpha) = \alpha C_1(t) + \frac{1}{2} \alpha^2 \dot{C}_1(t) + \dots$ , and so (1) can be interpreted as a first-order approximation.

sumer solves the problem

$$(2) \quad \max \int_0^{\infty} e^{-\rho t} U(C_1, C_2) dt$$

subject to  $M \geq \alpha C_1$ ,  $M^* \geq \alpha C_2$ ,

$$\dot{A} = L - C_1 - C_2, \quad A(0) = A_0,$$

where the initial value of assets  $A_0 = M_0 + M_0^*$  is assumed greater than  $\alpha L$  so the value of consumption can exceed labor income, and  $\rho$  is the discount rate.

Before formally solving this problem, an intuitive treatment is helpful. Suppose the consumer has made an optimal savings decision (i.e., chosen  $\dot{A}$ ), and then faces the budget constraint  $C_1 + C_2 = L - \dot{A}$  shown as  $BB$  in Figure 1. With a constant exchange rate of unity, and instantaneous exchanges between domestic and foreign money permitted at the central bank, the constraints (1) can be combined into

$$(1') \quad A \geq \alpha(C_1 + C_2).$$

If at a point in time the consumer has large asset balances, then the transactions constraint (1') may be illustrated as  $RR$  in Figure 1, which is not binding given the budget constraint  $BB$ . In this case, consumption would occur at  $P$ , where the marginal rate of substitution along the indifference curve  $U$  equals the ratio of international prices.

Alternatively, with small money balances the transactions constraint may be illustrated by  $R'R'$  in Figure 1, lying inside the budget constraint  $BB$ . In this case consumption would occur at  $P'$ , and the consumer would experience "forced savings" due to an inability to satisfy the transactions constraint at the desired consumption level. It is important to notice that at  $P'$ , the marginal rate of substitution still equals the ratio of international prices, since the budget and transactions constraints are parallel.<sup>4</sup>

<sup>4</sup>The budget and transactions constraints are parallel since I have assumed the same value of  $\alpha$  in the domestic and foreign transactions constraints (1). If different values of  $\alpha$  were used in the two constraints in (1), then this would imply a further reason for using direct trade controls to support the social optimum.

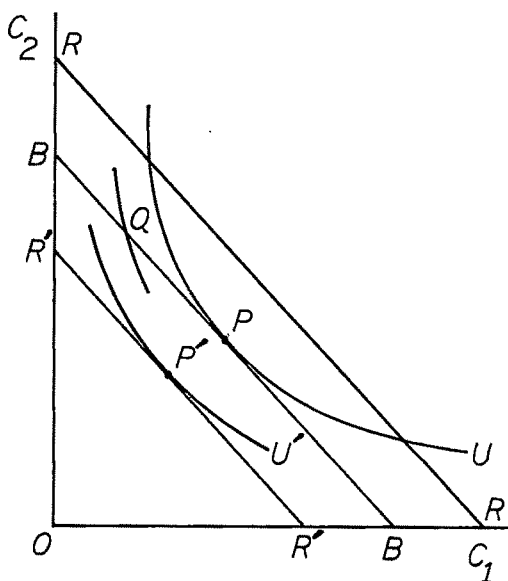


FIGURE 1

Turning to a formal solution of problem (2), when the consumer is *liquid* (i.e.,  $M > \alpha C_1$  or  $M^* > \alpha C_2$ ), the optimality conditions are

$$(3) \quad U_1 = U_2 = \theta, \quad \dot{\theta} = \rho\theta,$$

where  $U_i = \partial U / \partial C_i$  for  $i=1,2$ , and  $\theta$  is interpreted as the shadow value of wealth.<sup>5</sup>

When the consumer is *illiquid* (i.e.,  $M = \alpha C_1$  and  $M^* = \alpha C_2$ ), then the optimality conditions can be written as

$$(4) \quad U_1 = U_2 = \theta + \alpha\phi, \quad \dot{\theta} = \rho\theta - \phi,$$

where  $\phi \geq 0$  is a Lagrange multiplier that is positive when the transactions constraints (1) are binding. Note that the optimality conditions still ensure that  $U_1/U_2=1$ , so the marginal rate of substitution equals the ratio of international prices. Differentiating the

<sup>5</sup>These conditions are obtained from the current value Hamiltonian  $H = U(C_1, C_2) + \lambda_1(M - \alpha C_1) + \lambda_2(M^* - \alpha C_2) + \phi(A - M - M^*) + \theta(L - C_1 - C_2)$ , with the optimality conditions  $\partial H / \partial C_i = \partial H / \partial M = \partial H / \partial M^* = 0$  and  $\dot{\theta} = \rho\theta - \partial H / \partial A$ ,  $i=1,2$  ( $\theta$  can be interpreted as the shadow value of wealth). When the consumer is liquid then  $\lambda_1 = \lambda_2 = \phi = 0$ .

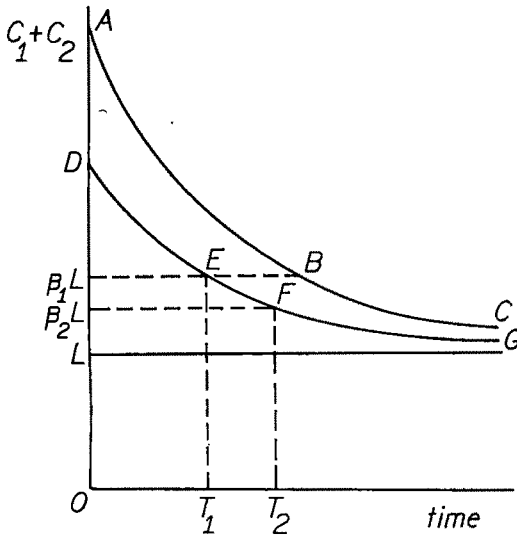


FIGURE 2

binding transactions constraint and using  $A = L - C_1 - C_2$ , we obtain

$$(5) \quad (\dot{C}_1 + \dot{C}_2)/(C_1 + C_2) = -(1/\alpha)[1 - L/(C_1 + C_2)].$$

For convenience let us assume that the instantaneous utility function is homogeneous of degree  $(1 - \sigma)$ ,  $0 < \sigma < 1$ . It follows from (3) that when the consumer is liquid,  $C_1$  and  $C_2$  both fall at the exponential rate  $\rho/\sigma$ . Also, note that in the long run the value of consumption  $(C_1 + C_2)$  will converge to wage earnings  $L$ , since in the absence of capital markets the savings obtained by foregoing current consumption and choosing  $(C_1 + C_2) < L$  could only generate extra consumption in the future, which would be suboptimal. The remaining question, then, is to determine when the transactions constraints are binding so the consumer is illiquid.

**LEMMA 1:** *The consumer is liquid (illiquid) according as  $C_1 + C_2 > (<) \beta_1 L$ , where  $\beta_1 \equiv 1/(1 - \alpha\rho/\sigma) > 1$ .*

The proof of this lemma rests upon setting (5) equal to  $-\rho/\sigma$ , which yields the point at which the consumer shifts from being liquid to illiquid;  $\alpha\rho/\sigma < 1$  is assumed.

The economy's consumption path is shown by  $ABC$  in Figure 2. Along  $AB$  the consumption of both goods falls at the exponential rate  $\rho/\sigma$ . The range over which the consumer is illiquid is shown by  $BC$ , where the value of consumption falls more slowly than  $\rho/\sigma$ , as indicated by (5). At every point of time, consumption exceeds labor income, so the economy has a balance of trade deficit. This deficit is financed by a reduction in the consumer's cash balances, which leads to a demand for central bank reserves whenever the consumer switches from domestic to foreign money to purchase the import. In the following section I discuss methods to obtain balance of payments equilibrium in the economy.

## II. Balance of Payments Equilibrium

With consumption exceeding labor income along the path  $ABC$  in Figure 2, the economy has a balance of trade deficit. The accumulated long-run deficit is given by the reduction in asset holdings (i.e.,  $A_0 - \alpha L$ , where  $\alpha L$  are the long-run asset holdings of the consumer as  $(C_1 + C_2)$  approaches  $L$ ). The total demand for central bank foreign exchange is then given by the integrated balance of trade deficit less the reduction in the consumer's foreign asset holdings, that is,

$$(6) \quad (A_0 - \alpha L) - (M_0^* - \alpha\gamma_2 L) = M_0 - \alpha\gamma_1 L,$$

where  $\gamma_i$  is the marginal and average propensity to consume good  $i$ , so  $\alpha\gamma_i L$  is the consumer's long-run holdings of domestic ( $i = 1$ ) and foreign ( $i = 2$ ) money.

If we now assume that initial central bank reserves of foreign exchange, denoted by  $R_0^*$ , are less than the accumulated private demand for foreign exchange given in (6), then the economy faces an unsustainable balance of trade deficit. Let us suppose that the central bank reserves include all available foreign borrowing (assumed to occur at a zero interest rate for simplicity). Several methods can then be considered to accommodate the trade deficit.

### A. Monetary Policy and Direct Trade Controls

First, a reduction in consumer holdings of domestic money will reduce the demand for foreign exchange. From (6), the level of initial money holdings needed to ensure that foreign exchange demand can be financed by available reserves is simply  $M'_0 = R^*_0 + \alpha\gamma_1 L$ . Second, an unanticipated devaluation would lower the foreign exchange value of domestic money holdings, and therefore have the same impact as directly reducing domestic balances.

Third, an equivalent way to restore balance of payments equilibrium is to impose an equal import tariff and export subsidy, thereby raising domestic prices. The higher domestic prices will have a real balance effect on consumption expenditure and thus reduce the balance of trade deficit. This uniform tariff-cum-subsidy policy ensures that the consumer's marginal rate of substitution equals the ratio of international prices: in the following section I shall examine the social optimality of such a policy.

The first two methods discussed above—reducing domestic money holdings and the unanticipated devaluation—may not be available to government authorities in a less developed country. In the absence of substantial capital markets, the central bank cannot rely on an open market operation to reduce money holdings, and in practice may attempt a recoinage of the domestic currency. This policy is essentially equivalent to the unanticipated devaluation that could not be used repeatedly. Accordingly, I shall next examine a devaluation which is anticipated.

### B. Anticipated Devaluation

Let us consider a devaluation which is *anticipated* from  $t = 0$ . Suppose that at time  $T'$  the central bank will devalue the domestic currency, raising the official price of foreign exchange from unity to  $x' > 1$ . As in Paul Krugman (1979) and Obstfeld (1984), I wish to solve for the perfect foresight path of the exchange rate between time 0 and  $T'$ . I shall impose exchange controls when the central bank is drained of foreign exchange

reserves, so that no more purchases can be made.<sup>6</sup> In that case a black market for foreign exchange is created with a floating exchange rate. Arbitrage and perfect foresight will ensure that the floating exchange rate is continuous after time 0, though it can jump initially, and equals the official rate  $x'$  at time  $T'$ .

The path of the floating exchange rate is solved from the conditions of portfolio equilibrium for the consumer. When the consumer is liquid, the optimality conditions are given by<sup>7</sup>

$$(7) \quad U_1 = \theta(1 + \alpha\dot{x}/x),$$

$$U_2 = \theta, \quad \dot{\theta} = \rho\theta,$$

which reduce to (3) when  $\dot{x} = 0$ . So long as the exchange rate is depreciating along this path, the consumer will not hold any domestic money balances in excess of those needed for consumption, and so  $M/x = \alpha C_1$ . If the central bank's reserves have been drained, then the consumer's holdings of domestic money must be  $M_0 - R^*_0$ , that is, initial holdings less the sale of domestic money to the central bank at the exchange rate of unity. Substituting these relationships into (7) we obtain

$$(8) \quad U_1[(M_0 - R^*_0)/\alpha x, C_2] = \theta(1 + \alpha\dot{x}/x)$$

$$U_2[(M_0 - R^*_0)/\alpha x, C_2] = \theta, \quad \dot{\theta} = \rho\theta.$$

<sup>6</sup>In Section IV, I shall permit the central bank to retain some amount of reserves before exchange controls are imposed, and gradually deplete these reserves in supporting the social optimum (see fn. 11). For simplicity, I do not examine this case now. Obstfeld also considers the case where exchange controls are imposed before reserves are drained, and Jorge de Macedo (1982) examines exchange rate behavior under currency inconvertibility.

<sup>7</sup>Assets are now given by  $A = M/x + M^*$  and so  $\dot{A} = \dot{M}/x + \dot{M}^* - (M/x)(\dot{x}/x) = L - C_1 - C_2 - (M/x)(\dot{x}/x)$ . It follows that the current value Hamiltonian is given by  $F = U(C_1, C_2) + \lambda_1(M/x - \alpha C_1) + \lambda_2(M^* - \alpha C_2) + \phi(A - M/x - M^*) + \theta\dot{A}$  with the optimality conditions  $\partial F/\partial C_i = \partial H/\partial M = \partial H/\partial M^* = 0$  and  $\dot{\theta} = \rho\theta - \partial H/\partial A$ ,  $i = 1, 2$ . When the consumer is liquid, then  $\lambda_2 = \phi = 0$ .



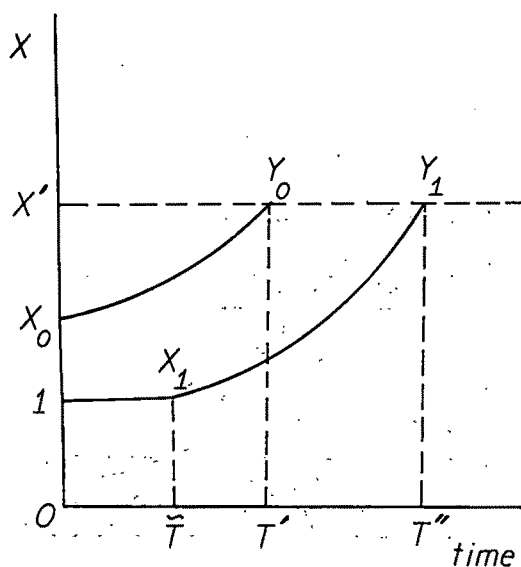


FIGURE 3

By inspection, the path of the floating exchange rate is solved from (8) as

$$(9) \quad x = x_0 e^{(\rho/\sigma)t}$$

To see this, with the exchange rate depreciating at the rate  $\rho/\sigma$ , then  $C_2$  also falls at this exponential rate in (8) and the marginal utilities rise at the rate  $\rho$ , as required. The initial value of the exchange rate is simply obtained by setting  $x(T') = x'$  and then inverting (9) to obtain

$$(10) \quad x_0 = x' e^{-(\rho/\sigma)T'}$$

If the initial value of the exchange rate  $x_0$  from (10) exceeds unity, indicating a discrete depreciation, then the central bank reserves are completely drained at time 0 and a black market for foreign exchange is created. In this case the path of the floating exchange rate is shown by  $X_0Y_0$  in Figure 3. Along this path the exchange rate continuously depreciates at the exponential rate  $\rho/\sigma$ , and equals the official rate at the time  $T'$ .<sup>8</sup>

<sup>8</sup>It appears that we cannot rule out a second equilibrium path, in which the exchange rate jumps to  $x'$

A slightly different path is obtained if knowledge of the devaluation *does not* drain the central bank's reserves at  $t = 0$ . Thus, in Figure 3 suppose the official devaluation will occur at time  $T''$ , where the value of  $x_0$  obtained from (12) indicates a discrete *appreciation* initially. However, it is impossible for the black market rate to appreciate when the central bank is prepared to sell domestic currency at the price of unity. It follows that the exchange rate is constant at unity until the time  $\bar{T}$ , when central bank reserves are exhausted.

Up to  $\bar{T}$  the consumption of both goods falls at the exponential rate  $\rho/\sigma$  (assuming the consumer is liquid, as described in Lemma 1). Falling demand for the exportable implies that domestic currency requirements to satisfy the transactions constraint are also falling, and so this domestic money is used to purchase foreign exchange. At time  $\bar{T}$  the central bank reserves are drained, and currency trading in the black market begins to operate. After this time along the path  $X_1Y_1$ , the floating exchange rate depreciates at the rate  $\rho/\sigma$ , and equals the official rate at  $T''$ .

After the official devaluation has occurred (at  $T'$  or  $T''$ ), the central bank has a zero stock of reserves, but the official market still finances the flow demand for foreign exchange. That is, the official rate  $x'$  is chosen so that the supply of foreign exchange from exports equals the value of imports minus the reduction in the consumer's foreign exchange holdings. This choice of  $x'$  achieves balance of payments equilibrium, as will be illustrated in Section IV. In the long run, the consumer's foreign exchange holdings approach  $\alpha\gamma_2L$ , and the flow supply of foreign exchange from exports just equals the value of imports.

immediately. The multiple equilibria occur since, unlike the models of Krugman and Obstfeld, the allocation of assets beyond those needed for transactions is indeterminate when the expected devaluation is zero. Since I wish to illustrate the possible inefficiency which arises from currency flight, I shall focus on the equilibrium path with a depreciating black market rate.

An important question is whether the floating exchange rate path implied by the currency fight and resulting black market imposes any welfare cost on the economy. Intuitively, one expects that the depreciating exchange rate will lead the consumer to forgo some consumption of the exportable and therefore reduce holdings of domestic currency required for transactions. With  $\alpha$  units of domestic money needed per unit of the exportable and the exchange rate depreciating at the rate  $\rho/\sigma$ , the relevant price of the exportable becomes  $(1 + \alpha\rho/\sigma)$ . As is verified from (7) and (9), the consumer's marginal rate of substitution equals  $(1 + \alpha\rho/\sigma)$  along the floating exchange rate path. In Figure 1, consumption occurs at a point such as  $Q$ , with less consumption of the exportable and greater consumption of the importable than at a point of tangency between the indifference curve and budget constraint.

In contrast the point  $P$ , where the marginal rate of substitution equals the ratio of international prices, is obtained under the uniform tariff-cum-subsidy policy discussed above. It is clear that to evaluate the merits of direct trade controls over the anticipated devaluation we must first solve for the social optimum. This is done in the next section, where I consider the optimum obtained by a central planner with full control over all variables. Section IV then examines various means to decentralize the social optimum.

### III. Social Optimum

A social planner faces the same problem as the representative consumer, but has the ability to print any amount of domestic money. It follows that the transactions constraint  $M \geq \alpha C_1$  will never be binding, and can be ignored. In addition, the social planner has access to all foreign exchange in the economy (i.e., that of the consumer ( $M^*$ ) and central bank ( $R^*$ )). Denoting these foreign exchange holdings by  $F^* \equiv M^* + R^*$ , the social planner solves

$$(11) \quad \max \int_0^\infty e^{-\rho t} U(C_1, C_2) dt$$

subject to  $F^* \geq \alpha C_2$ ,  $\dot{F}^* = L - C_1 - C_2$ ,

$$F^*(0) = M_0^* + R_0^*,$$

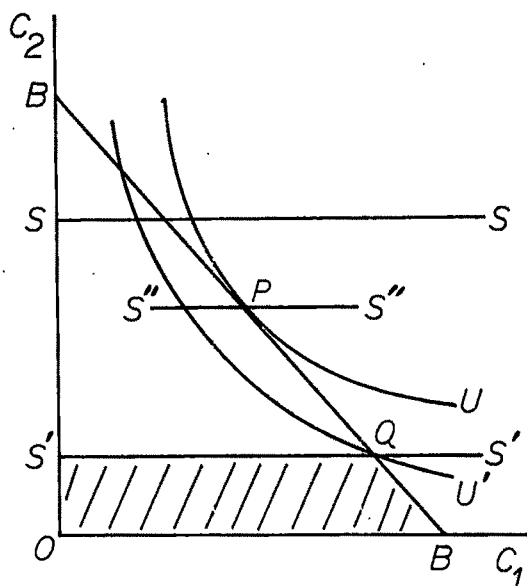


FIGURE 4

where  $(L - C_1)$  are exports of good 1,  $C_2$  are imports of good 2, and so  $(L - C_1 - C_2)$  is the balance of trade deficit and flow reduction in foreign exchange.

It should be noted that while the constraint  $M \geq \alpha C_1$  in (11) is ignored, when we decentralize the solution we shall never consider policies in which a large (or any) amount of domestic money is printed. Instead, a zero shadow value on the constraint  $M \geq \alpha C_1$ , which applies for the social planner, shall be achieved in the private economy through direct trade controls. In addition, in the decentralized solution the central bank will not have access to the private holdings  $M^*$  of foreign exchange.

Before formally solving this problem, an intuitive treatment is again useful. Suppose the social planner has made an optimal savings decision (i.e., chosen  $\dot{F}^*$ ), and then faces the budget constraint  $C_1 + C_2 = L - \dot{F}^*$  shown as  $BB$  in Figure 4. If at a point in time the foreign exchange holdings  $F^*$  are large, the transactions constraint  $F^* \geq \alpha C_2$  may be illustrated as  $SS$  which is not binding. In this case consumption would occur at  $P$ , and the social marginal rate of substitution along the indifference curve  $U$  would equal the ratio of international prices.

However, if the economy's foreign exchange holdings are small, then the transactions constraint may be illustrated by  $S'S'$  in Figure 4, and the available consumption set for the social planner is the shaded polygon  $OS'QB$ . The optimal consumption point would be  $Q$ , and the social marginal rate of substitution would differ from international prices. At the point  $Q$ , less of the importable good ( $C_2$ ) is consumed than in the absence of the transactions constraint on foreign exchange.

I shall now show that it is *optimal* for the social planner to deplete foreign exchange holdings to the point such as  $Q$ , where the relative marginal utility of consuming the importable exceeds its international price. This result is intuitively understood by first considering the transactions constraint  $S'S'$  in Figure 4, which would just permit the economy to consume at  $P$ . A slight reduction in foreign exchange holdings beyond this point would bring a gain from consumption of the importable. However, the loss to the economy by forcing the marginal rate of substitution to differ from international prices is initially of the second-order of smalls. Thus, starting at  $P$  the social planner would choose to deplete foreign exchange holdings further, moving the economy to a point such as  $Q$ . This behavior is formally analyzed as follows.

The optimality conditions for problem (11) are given by

$$(12) \quad U_1 = \mu, \quad U_2 = \mu + \alpha\lambda, \quad \dot{\mu} = \rho\mu - \lambda,$$

where  $\mu$  is the shadow price of foreign exchange and  $\lambda \geq 0$  is the Lagrange multiplier for the constraint  $F^* \geq \alpha C_2$ .<sup>9</sup> If this foreign exchange constraint is not binding, then  $\lambda = 0$  and conditions (12) imply that consumption of both goods falls at the exponential rate  $\rho/\sigma$ , with  $U_1 = U_2$ . When the foreign exchange constraint is binding, however, then  $\lambda > 0$  so  $U_2 > U_1$ , indicating that the social planner shifts consumption towards the export good 1 to economize on imports and

foreign exchange requirements. Indeed, in the long run,  $\dot{\mu} = 0$ , so  $\lambda = \rho\mu$  and  $U_2/U_1 = 1 + \alpha\rho$ , indicating a long-run divergence between the marginal rate of substitution and international price ratio.

The long-run formula  $U_2/U_1 = 1 + \alpha\rho$  can be readily understood. The marginal benefit from consuming a unit of the importable good 2, relative to good 1, is  $U_2/U_1$ . The marginal cost is its price of unity *plus* the cost of accumulating foreign exchange needed to satisfy the transactions constraint  $F^* \geq \alpha C_2$ . To consume one unit of the importable good,  $\alpha$  units of foreign exchange must be accumulated, which is possible by delaying consumption of the export good. The cost of delaying consumption is simply the discount rate. Thus, the marginal benefit of consuming the importable good equals its marginal cost of  $1 + \alpha\rho$ .

Next, it should be determined when the foreign exchange constraint  $F^* \geq \alpha C_2$  is binding:

LEMMA 2: *In the social optimum, the foreign exchange constraint is (not) binding according as  $C_1 + C_2 > (<) \beta_2 L$ , where  $\beta_2 \equiv 1/(1 - \gamma_2 \alpha \rho / \sigma)$ , and  $\gamma_2$  is the marginal and average propensity to consume good 2.*

The proof of this lemma follows by noting that when the foreign exchange constraint is binding we have  $\dot{F}^* = \alpha \dot{C}_2 = L - C_1 - C_2$ , and so

$$(13) \quad \dot{C}_2/C_2 = -(1/\alpha\gamma_2)[1 - L/(C_1 + C_2)]$$

The first point at which the foreign exchange constraint becomes binding is found by setting (13) equal to  $-\rho/\sigma$ , which is the rate of change of consumption without the constraint. Lemma 2 is thus obtained. Let us assume that  $0 < \gamma_2 < 1$ , and it follows that  $\beta_1 > \beta_2 > 1$ .

The socially optimal consumption path is shown as  $DEFG$  in Figure 2. I shall henceforth let  $T_1$  denote the time at which the consumer becomes illiquid (i.e.,  $(C_1 + C_2) = \beta_1 L$ ) along the optimal consumption path, and let  $T_2$  denote the time at which the social foreign exchange constraint  $F^* \geq \alpha C_2$  becomes binding along this path. Between times 0 and

<sup>9</sup>The Hamiltonian is  $H = U(C_1, C_2) + \lambda(F^* - \alpha C_2) + \mu(L - C_1 - C_2)$ , with optimality conditions  $\partial H / \partial C_i = 0$ ,  $i = 1, 2$ , and  $\dot{\mu} = \rho\mu - \partial H / \partial F^*$ .

$T_2$  the optimal consumption of both goods falls at the exponential rate  $\rho/\sigma$ , along the path  $DF$  in Figure 2. After  $T_2$  consumption,  $C_2$  falls at the slower rate given by (13), along the path  $FG$ . After the foreign exchange constraint is binding the relative marginal utility of consuming the importable exceeds its international price. In the following section I shall investigate methods to decentralize this social optimum.

#### IV. Policies to Obtain the Optimum

Let us consider several properties of the social optimum. For times beyond  $T_2$  in Figure 2, the social foreign exchange constraint is binding so that  $F^* = \alpha C_2$ . However, from the consumer's problem solved in Section I we also know that  $M^* = \alpha C_2$  for times beyond  $T_1$ . Since total foreign exchange  $F^*$  is defined as private holdings  $M^*$  plus reserves  $R^*$ , it follows that *government reserves are exhausted at time  $T_2$  and beyond*. That is, any policy leading to the social optimum will allow government reserves to be run down to zero at time  $T_2$ , and afterwards the official market will simply finance flow transactions in foreign exchange. The value of consumption exceeds labor income at time  $T_2$  and beyond, implying a balance of trade deficit, and the foreign exchange needed to finance imports is obtained from exports plus a reduction in the consumer's holdings of  $M^*$ .

Whenever the social foreign exchange constraint is not binding (i.e.,  $F^* > \alpha C_2$ ), then the marginal rate of substitution should equal the ratio of international prices to obtain the optimum. In this case, equal import tariffs and export subsidies could be used to reduce consumption through a real balance effect. When the social foreign exchange constraint is binding, however, the import tariff should exceed the export subsidy. The level of tariff and subsidy needed to reduce consumption to the socially optimal level also depends on the exchange rate. If the domestic currency is devalued through a discrete anticipated change or a crawling peg, then the required level of trade controls is lower.

The social optimum can be decentralized through a variety of policies, which I now examine in detail.

#### A. Direct Trade Controls

Let  $p_i$  denote the domestic price of good  $i$ ,  $i = 1, 2$ , so  $(p_1 - 1)$  is the export subsidy and  $(p_2 - 1)$  is the import tariff. Let us assume that tariffs on imported items are paid in the domestic currency. Then the transactions constraints (1) become

$$(14) \quad M \geq \alpha p_1 C_1 + \alpha (p_2 - 1) C_2,$$

$$M^* \geq \alpha C_2.$$

The difference between tariff revenues and subsidy payments are redistributed to (or collected from) the consumer as lump sum transfers. It is clear from (14) that raising the export subsidy or import tariff increases the amount of domestic currency needed to satisfy the transactions constraints, and therefore prevents this wealth from being spent on future consumption. In this way, raising the tariff or subsidy exerts a real balance effect on consumption.

The pattern of direct trade controls needed to support the social optimum is summarized as follows (the times  $T_1$  and  $T_2$  are defined following Lemma 2, and shown in Figure 2):

**PROPOSITION 1:** *The social optimum can be achieved by an import tariff and export subsidy satisfying:*

- (a) *for  $0 \leq t \leq T_1$  the tariff and subsidy are equal and constant;*
- (b) *for  $T_1 < t \leq T_2$  the tariff and subsidy are equal and rise gradually so that consumption continues to fall at the exponential rate  $\rho/\sigma$ ;*
- (c) *at  $T_2$  the tariff and subsidy are at a level that ensures that government reserves are exhausted;*
- (d) *for  $t > T_2$  the tariff rises above the subsidy, and in the long run, the price ratio approaches  $p_2/p_1 = 1 + \alpha p$ .*

Proposition 1 is explained as follows.<sup>10</sup> Over the range  $0 \leq t \leq T_1$  the private and socially optimal consumption of both goods falls at the rate  $\rho/\sigma$ . Then the tariff and subsidy are only used to exert a real balance effect on consumption. For  $T_1 < t \leq T_2$  the

<sup>10</sup> Formal proofs of this and other propositions are contained in my working paper, available on request.

consumer is illiquid, so private consumption would begin to fall at a rate slower than  $\rho/\sigma$  (as described following Lemma 1). However, the social foreign exchange constraint is not yet binding, so the government should intervene with a *rising* tariff and subsidy to ensure that consumption continues to fall at the rate  $\rho/\sigma$ . This range illustrates that government intervention is needed when the private transactions constraints are binding but the social constraint is not.

At time  $T_2$  the equal tariff and subsidy, denoted by  $p$ , will satisfy

$$(15) \quad M_0 - R_0^* = \alpha p C_1 + \alpha(p-1)C_2,$$

which states that domestic money demand equals  $M_0 - R_0^*$ . Thus, the reduction of the consumer's domestic money stock from  $M_0$  to this level implies that exactly  $R_0^*$  foreign exchange has been purchased from the central bank, depleting reserves. After time  $T_2$  the social foreign exchange constraint is binding, implying that the social planner shifts consumption towards the exportable to economize on imports and foreign exchange requirements. This shift in consumption is decentralized through the import tariff rising above the export subsidy. In the previous section it was found that the socially optimal ratio of marginal utilities approaches  $1 + \alpha\rho$  in the long run, implying the same ratio of domestic prices.

The pattern of trade controls described by Proposition 1 are numerically illustrated in Table 1. A discount rate of 5 percent per year is used, or 0.41 percent per month. From the *World Development Report 1981* (Table 15) the average ratio of international reserves to monthly imports for low-income countries is 2.8 in 1979, so  $\alpha = 2.8$  is chosen.<sup>11</sup> The marginal and average propensity to spend on the importable ( $\gamma_2$ ) is set at a relatively high value of 0.5. A Cobb-Douglas

utility function is used, and I consider the cases where the utility function is homogeneous of degree 0.1 ( $\sigma = 0.9$ ) and degree 0.9 ( $\sigma = 0.1$ ). In the latter case, the utility function nearly exhibits constant returns to scale, and assets are spent on consumption at a rapid rate.<sup>12</sup>

Let us suppose that initial holdings of domestic money are  $M_0 = 300$  while  $L = 100$ , so under *laissez-faire* the consumer would demand  $M_0 - \alpha\gamma_1 L = 160$  units of foreign exchange from the central bank. However, reserves are only one-half as much,  $R_0^* = 80$ . We can compute that the unanticipated devaluation needed to restore balance of payments equilibrium is  $x = 1.36$  or 36 percent.<sup>13</sup> Let us also suppose that the initial consumer holdings of foreign money are  $M_0^* = 300$ , so the total foreign exchange available to a social planner is  $F_0^* = 380$ . This information permits us to compute the times  $T_1$  and  $T_2$  along the social optimum, for varying values of  $\sigma$ .

The direct trade controls needed to obtain the social optimum are reported in Table 1, Part A. The optimum is achieved by initially setting an import tariff and export subsidy of 27.8 percent, with  $\sigma = 0.9$ . This constant level is held until the consumer becomes illiquid at  $T_1$ , and is then gradually increased to reach 28.1 percent at  $T_2$ . After this time the tariff rate rises above the subsidy, and in the long run their ratio is  $1 + \alpha\rho = 1.0115$  with the assumed parameter values. For other reasonable values of  $\alpha$  and  $\rho$  the difference between the long-run rates would still be small. For  $\sigma = 0.1$  the times  $T_1$  and  $T_2$  are recomputed, leading to different values of the trade controls along the adjustment path. However, the qualitative behavior of the trade controls in this case is the same as before:

<sup>11</sup>This calculation of  $\alpha$  is justified by interpreting  $M^*$  as claims by the representative consumer on foreign exchange reserves of the central bank. Also, McKinnon and Donald Mathieson (1981, p. 5) report that the ratio of domestic  $M_2$  to monthly GNP is 2.4 for typical Latin American economies, and 2.88 for Asia. So  $\alpha = 2.8$  appears consistent with the domestic transactions constraint, as well as the foreign.

<sup>12</sup>The value  $\sigma = 0.1$  is somewhat smaller than is supported by empirical evidence. Thus, the parameter  $-\sigma$  can be interpreted as the "expenditure elasticity of the marginal utility of income," and Ragnar Frisch (1959) conjectured it would lie in the range  $(-10, -0.1)$  with lower values applying to poorer countries. Subsequent estimates summarized in Alan Brown and Angus Deaton (1972, p. 1206) have established a range  $(-5, -0.5)$ , while in my model,  $-\sigma$  is restricted to the range  $(-1, 0)$ .

<sup>13</sup>This exchange rate is computed from the formula  $(M_0/x) - \alpha\gamma_1 L = R_0^*$ .

TABLE 1

$\sigma$		Initial Values	Values at $T_1$	Values at $T_2$	Long-Run Values	$T_1$ (months)	$T_2$
<b>A. Direct Trade Controls<sup>a</sup></b>							
0.9	Import						
	Tariff	27.8	27.8	28.1	29.0	28.8	30.2
	Export						
	Subsidy	27.8	27.8	28.1	27.6		
0.1	Import						
	Tariff	21.0	21.0	24.0	29.0	7.0	8.6
	Export						
	Subsidy	21.0	21.0	24.0	27.6		
<b>B. Anticipated Devaluation and Direct Trade Controls<sup>a</sup></b>							
0.9	Exchange						
	Rate	1.33	1.52	1.52	1.52	28.8	30.2
	Import						
	Tariff	1.26	1.28	1.45	2.08		
	Export						
	Subsidy	0	1.28	1.45	0.92		
0.1	Exchange						
	Rate	1.0	1.16	1.16	1.16	7.0	8.6
	Import						
	Tariff	0 <sup>b</sup>	11.5	13.8	18.1		
	Export						
	Subsidy	0	11.5	13.8	16.7		

<sup>a</sup>The tariff and subsidy rates are shown in percent.

<sup>b</sup>In this case foreign exchange reserves of the central bank are not drained until after 3.6 months, at which time an import tariff of 10.1 percent is applied.

the tariff and subsidy are constant and equal between times 0 and  $T_1$ , rise gradually to  $T_2$ , and after this point the tariff exceeds the subsidy to reflect the binding social foreign exchange constraint. The long-run optimal rates for the tariff and subsidy are independent of  $\sigma$ .

#### B. Anticipated Devaluation and Direct Trade Controls

A second method to obtain the social optimum is through a devaluation that is anticipated, combined with trade controls to correct the distorting effect of the currency flight on consumption. Thus, in Section II.B I argued that prior to the devaluation the central bank reserves would be drained and a floating, black market exchange rate established. This floating rate depreciates at the exponential rate  $\rho/\sigma$ , which implies that the consumer holds less domestic money and decreases consumption of the exportable to where  $U_1/U_2 = 1 + \alpha\rho/\sigma$ . To correct this dis-

tortion in consumption and achieve the social optimum, the following policy can be used:

**PROPOSITION 2:** *The social optimum can be achieved by an anticipated devaluation at time  $T' = T_1$  and trade controls satisfying:*

(a) *during the period before the devaluation, when central bank reserves are drained, an import tariff of  $(\alpha\rho/\sigma)/(1 + \alpha\rho/\sigma)$  is used;*

(b) *at  $T' = T_1$  the import tariff and export subsidy are both set at  $\alpha\rho/\sigma$ , and remain equal while rising until time  $T_2$ ;*

(c) *for  $t > T_2$  the tariff and subsidy are used as described in Proposition 1.*

Proposition 2 is explained as follows.<sup>14</sup> When the black market exchange rate is de-

<sup>14</sup>I have chosen to have the official devaluation occur at time  $T_1$  for convenience. This timing implies that the central bank will have to satisfy some demand for

preciating, the consumer shifts asset holdings towards foreign exchange, and the import tariff described in (a) is needed to offset the incentive to consume the foreign good. Immediately after the devaluation occurs, the tariff and subsidy should be equalized, but they cannot be set at zero for the following reason. To ensure that consumption follows the social optimum for  $0 \leq t \leq T_1$ , the relevant "price" of each good should be constant. Since  $\alpha$  units of domestic money must be held per unit consumption of good 1 its price equals  $p_1$  plus the loss through exchange rate depreciation, giving  $\pi = p_1(1 + \alpha \dot{x}/x)$  as the magnitude which must be constant. Thus, when the black market depreciation is reduced from  $\alpha\rho/\sigma$  to zero at the official devaluation, the nominal price of good 1 is correspondingly increased so that  $\pi$  is constant. To ensure that the consumption path chosen by the consumer coincides with the social optimum for  $T_1 < t \leq T_2$ , the tariff and subsidy should rise as described in Proposition 1. Beyond  $T_2$  the tariff exceeds the subsidy, and in the long run their ratio is  $p_2/p_1 = (1 + \alpha\rho)$ .

The appropriate devaluation and trade controls are numerically illustrated in Table 1, Part B, where an economy is considered with the same parameter values as in Section IV.A. The unanticipated devaluation needed to restore balance of payments equilibrium is 36 percent. With  $\sigma = 0.9$ , central bank reserves are drained immediately and the black market exchange rate jumps from unity to 1.33, thereafter depreciating at the exponential rate  $\rho/\sigma = 0.46$  percent. An import tariff of 1.26 percent is needed to offset the incentive to hoard foreign exchange and shift consumption towards the importable. At time  $T' = T_1$  the black market rate equals the official exchange rate of 1.52, which then remain constant. The official devaluation of

52 percent exceeds the unanticipated devaluation since consumers are now holding a minimum amount of domestic money. It follows that a greater devaluation is needed to achieve balance of payments equilibrium. At  $T_1$  the tariff and subsidy are adjusted slightly to 1.28 percent and rise to 1.45 percent at time  $T_2$ ; beyond this point the tariff exceeds the subsidy.

With  $\sigma = 0.1$ , initial consumption of good 1 is higher than before, requiring larger holdings of domestic money and therefore less demand for foreign exchange. In this case, government reserves are drained only after consumption of good 1 and domestic money holdings have been suitably reduced, which occurs in 3.6 months. For  $0 \leq t \leq 3.6$  the tariff and subsidy are zero. When reserves are drained the black market exchange rate begins to depreciate at the rate  $\rho/\sigma = 4.1$  percent. An import tariff of 10.1 percent is then required to offset the distorting effect on consumption. At  $T_1$ , the black market rate equals the official exchange rate of 1.16, and the tariff-cum-subsidy are adjusted to 11.5 percent. For  $T_1 \leq t \leq T_2$ , the tariff and subsidies rates rise somewhat, and after  $T_2$ , the tariff exceeds the subsidy. The relatively high level of these trade controls imply that the official devaluation of 16 percent is sufficient to achieve balance of payments equilibrium.

The advantage of this policy over just using direct trade controls is that, in reality, trade controls may lead to some resource cost beyond that captured in our model. Anne Krueger (1974) has emphasized the substantial "rent-seeking" costs associated with trade and foreign exchange quotas, while Bhagwati and T. N. Srinivasan (1980) have suggested that these costs can apply to tariffs as well. In the presence of such costs, the anticipated devaluation-cum-trade control option would be preferred to only using trade controls. However, it seems plausible that in reality the currency flight caused by a devaluation may also cause greater inefficiency than captured in our model. This consideration leads us to consider a third option for obtaining the social optimum: direct trade controls which are gradually reduced using a crawling peg devaluation.

foreign exchange between  $T_1$  and  $T_2$ , and therefore withhold a minimal level of reserves before exchange controls are imposed. For the two cases shown in Table 1, the level of withheld reserves are 0.46 and 5.05 with  $\sigma$  equal to 0.9 and 0.1, respectively, as compared with initial central bank reserves of 80.

### C. Direct Trade Controls and Crawling Peg

With an exchange rate of  $x$ , the transactions constraints (16) become

$$(14') \quad M/x \geq \alpha p_1 C_1 + \alpha(p_2 - 1)C_2,$$

$$M^* \geq \alpha C_2,$$

where  $p_i$  is the foreign currency price of good  $i$ ,  $i=1,2$ , inclusive of the tariff and subsidy rates. It is clear from (14') that higher values of  $x$  obtained through exchange rate devaluation can be used to replace that tariff and subsidy, while keeping the transactions constraint binding. When the exchange rate devalues through a crawling peg, the consumer shifts asset holdings towards foreign money and consumes more of the imported good, as occurred under the black market devaluation. Offsetting trade controls are then needed to retain the social optimum, as indicated in the following policy:

**PROPOSITION 3:** *The social optimum can be achieved by a crawling peg devaluation beginning at time  $T_3 > T_2$  and direct trade controls satisfying:*

(a) *up to time  $T_3$  the tariff and subsidy are applied as in Proposition 1;*

(b) *for  $t \geq T_3$  a crawling peg is adopted and that the tariff rises further above the export subsidy. As the exchange rate devalues, the tariff and subsidy are correspondingly reduced.*

(c) *The crawling peg is eliminated when a chosen level  $\bar{x}$  has been reached and  $\dot{x}=0$ , after which the exchange rate is constant. By choosing  $\bar{x}$  suitably high the tariff and subsidy can be made as low as desired, subject to the long-run relation  $p_2/p_1 = 1 + \alpha\rho$ .*

In the proof of Proposition 3 I derive a family of paths that the exchange rate must follow to support the social optimum. The initial choice of  $\dot{x}(T_3)$  can be made by the government, and corresponding to each choice is a unique path for the exchange rate. Let  $\bar{x}$  denote the maximum value of the exchange rate obtained along this path, which occurs at  $\dot{x}=0$ . When  $\bar{x}$  is reached the crawling peg is eliminated. Then it can be

shown that  $\bar{x}$  can be made as high as desired by suitable choice of  $\dot{x}(T_3) > 0$ . It follows that the tariff and subsidy can be made as low as desired, subject to the long-run relation  $p_1/p_2 = 1 + \alpha\rho$ .

The policies described in Propositions 1, 2, and 3 each support the same social optimum, so the choice between them must rest on considerations outside the model. I argued above that direct controls may involve "rent-seeking" costs, and that currency flight may lead to significant inefficiency beyond that captured in the model. In that case the crawling peg option would be preferred, since currency flight is avoided while the trade controls used are temporary. The only drawback to the crawling peg option is that the tariff and subsidy would need to be frequently reduced to exactly achieve the social optimum, while in practice the level of such trade controls may be inflexible.

### V. Discussion and Conclusions

In a series of pioneering papers, Helpman and Razin have analyzed "cash-in-advance" constraints in a discrete time trade model. They find that with complete markets these constraints are neutral and do not cause any inefficiency; see especially Helpman (1981b) who shows that the Pareto optimum is preserved. In contrast, I have found three reasons for using trade controls:

1) an equal import tariff and export subsidy will lower consumption and reduce the balance of payments deficit without causing currency flight;

2) a depreciation causes consumers to shift asset holdings towards foreign money and therefore purchase more of the importable;

3) private consumers and the social planner face different transactions constraints.

The third reason above is due to *incomplete markets*, that is, I have assumed that this country's currency is not traded on international markets and also that consumers cannot shift wealth across time through bond markets. It follows that at times the consumer is illiquid, while at other times the



social planner is constrained in foreign exchange. Resolving the difference between these private and social constraints is a rationale for government intervention.

However, the first and second reasons listed above are not due to incomplete markets, but rather, are due to *real balance effects*. That is, a devaluation or uniform tariff-cum-subsidy will reduce the real value of domestic money and lower consumption, but the real balance effect of a devaluation can be avoided by holding the foreign currency. Thus, anticipated devaluations cause a nonneutral change in desired asset holdings and consumption patterns. A fundamental point that should be recognized is that the *transactions constraints in the basic model of Helpman and Razin are neutral because real balance effects are absent*.

That is, Helpman and Razin suppose that in the beginning of a period each consumer receives wages, dividends, and government transfers in cash, which are then allocated between purchases of bonds and consumption goods in that period or carried over to the next. It is shown that with positive nominal interest rates no cash balances are carried across periods (Helpman, 1981b; Helpman and Razin, 1982b). It follows that prices and the exchange rate can jump from one period to the next with no real effects, since the real balances of all consumers are zero. The only exception occurs with respect to *initial* holdings of domestic and foreign bonds, whose real values are affected by price movements in the first period. However, changes in the real values of initial assets only cause an income transfer across countries and do not result in inefficiency.

In modifications of the basic model, real balance effects occur and exchange rate movements are nonneutral. Thus, in Helpman and Razin (1982a), capital markets are incomplete and prices are uncertain, so money is held across periods and real balance effects occur. A precautionary role for money also arises in Stockman (1980) and Svensson, where anticipated devaluations have a nonneutral effect on consumption. Nonneutrality of the exchange rate regime is observed in David Aschauer and Jeremy Greenwood

(1983) where labor supply is endogenous and wage payments are lagged one period. A similar time pattern in the payments of dividends is assumed in Helpman and Razin (1984) and the earlier, autarky model of Stockman (1981), and in both cases inefficiency can arise depending on the nominal interest rate. Introducing real balance effects has been facilitated in my model through the use of continuous time cash-in-advance constraints, which ensure that money balances are strictly positive. As these examples make clear, real balance effects can be an important source of nonneutrality of exchange rate movements.

This paper differs in spirit, but should be complementary to, analyses of the inefficiency of exchange control regimes as in Bhagwati (1978) and Krueger (1978). Thus, recognizing that the extensive systems of quotas and trade restrictions that exist in many developing countries are nonoptimal, I have tried to determine why these countries do not simply devalue. The reason I have identified is that anticipated devaluations will cause currency flight, which imposes a welfare cost on society. Thus, to control the balance of payments and obtain the social optimum, a more sophisticated policy is needed, and I identified three such policies: direct trade controls which are permanently applied; an anticipated devaluation causing currency flight, combined with direct trade controls to offset the speculation; and direct trade controls which are gradually eliminated through a crawling peg devaluation.

Since all these policies support the same social optimum, the choice between them rests on considerations outside of our model. In many cases it seems that the third option would be preferred: a uniform tariff-cum-subsidy applied initially to control the balance of payment, and then gradually reduced by a crawling peg and adjusted to nonuniform (with the tariff exceeding the subsidy) reflecting the binding social foreign exchange constraint. Elements of this policy have been used in various countries, and overall it seems to be a considerable improvement over existing systems of exchange controls or the dictum to simply devalue.

## REFERENCES

- Aschauer, David and Greenwood, Jeremy, "A Further Exploration in the Theory of Exchange Rate Regimes," *Journal of Political Economy*, October 1983, 91, 868-75.
- Bhagwati, Jagdish N., "The Theory and Practice of Commercial Policy: Departures from Unified Exchange Rates," Special Papers in International Economics No. 8, International Finance Section, Princeton University, 1968.
- \_\_\_\_\_, "On the Equivalence of Tariffs and Quotas," in his *Trade, Tariffs, and Growth*, Cambridge: MIT Press, 1969, ch. 9.
- \_\_\_\_\_, *Anatomy and Consequences of Exchange Control Regimes*, Cambridge: Ballinger, 1978.
- \_\_\_\_\_, and Srinivasan, T. N., "Revenue Seeking: A Generalization of the Theory of Tariffs," *Journal of Political Economy*, December 1980, 88, 1069-87.
- Brown, Alan and Deaton, Angus, "Models of Consumer Behavior," *Economic Journal*, December 1972, 82, 1145-236.
- Caves, Richard E. and Jones, Ronald W., *World Trade and Payments*, 3rd ed., Boston: Little, Brown and Co., 1981.
- Clower, Robert W., "A Reconsideration of the Microfoundations of Monetary Theory," *Western Economic Journal*, December 1967, 6, 1-9.
- Frisch, Ragnar, "A Complete Scheme for Computing all Direct and Cross Demand Elasticities in a Model with Many Sectors," *Econometrica*, April 1959, 27, 177-96.
- Helpman, Elhanan, (1981a) "Inflation and Balance of Payments Adjustments with Maximizing Consumers," in M. June Flanders and Assaf Razin, eds. *Development in an Inflationary World*, New York: Academic Press, 1981.
- \_\_\_\_\_, (1981b) "An Exploration in the Theory of Exchange Control Regimes," *Journal of Political Economy*, October 1981, 89, 865-90.
- \_\_\_\_\_, and Razin, Assaf, "Towards a Consistent Comparison of Alternative Exchange Rate Regimes," *Canadian Journal of Economics*, August 1979, 12, 394-409.
- \_\_\_\_\_, and \_\_\_\_\_, (1982a) "A Comparison of Exchange Rate Regimes in the Presence of Imperfect Capital Markets," *International Economic Review*, June 1982, 23, 365-88.
- \_\_\_\_\_, and \_\_\_\_\_, (1982b) "Dynamics of a Floating Exchange Rate Regime," *Journal of Political Economy*, August 1982, 90, 728-54.
- \_\_\_\_\_, and \_\_\_\_\_, "The Role of Savings and Investment in Exchange Rate Determination under Alternative Monetary Mechanisms," *Journal of Monetary Economics*, May 1984, 13, 307-26.
- Kohn, Meir, "The Finance (Cash-in-Advance) Constraint Comes of Age: A Survey of Some Recent Developments in the Theory of Money," Working Paper No. 84-1, Dartmouth College, February 1984.
- Krueger, Anne O., "The Political Economy of the Rent-Seeking Society," *American Economic Review*, June 1974, 64, 291-303.
- \_\_\_\_\_, *Liberalization Attempts and Consequences*, Cambridge: Ballinger, 1978.
- Krugman, Paul R., "A Model of Balance-of-Payments Crisis," *Journal of Money, Credit and Banking*, August 1979, 11, 311-25.
- Lucas, Robert E., Jr., "Interest Rates and Currency Prices in a Two-Country World," *Journal of Monetary Economics*, November 1982, 10, 335-59.
- de Macedo, Jorge Braga, "Exchange Rate Behavior with Currency Inconvertibility," *Journal of International Economics*, February 1982, 12, 65-81.
- McKinnon, Ronald I., *Money and Capital in Economic Development*, Washington: The Brookings Institution, 1973.
- \_\_\_\_\_, and Mathieson, Donald J., "How to Manage a Repressed Economy," in *Essays in International Economics*, No. 145, International Finance Section, Princeton University, December 1981.
- Meier, Gerald M., *International Economics: The Theory of Policy*, New York: Oxford University Press, 1980.
- Obstfeld, Maurice, "Balance of Payments Crises and Devaluation," *Journal of Money, Credit and Banking*, May 1984, 16, 208-17.
- \_\_\_\_\_, and Stockman, Alan C., "Exchange Rate Dynamics," in Peter B. Kenen and Ronald W. Jones, eds., *Handbook of International Economics*, Amsterdam: North-Holland,

- 1984, ch. 17.
- Persson, Torsten, "Real Transfers in Fixed Exchange Rate Systems and the International Adjustment Mechanism," *Journal of Monetary Economics*, May 1984, 13, 349-70.
- Stockman, Alan C., "A Theory of Exchange Rate Determination," *Journal of Political Economy*, August 1980, 88, 673-98.
- \_\_\_\_\_, "Anticipated Inflation and the Capital Stock in a Cash-in-Advance Economy," *Journal of Monetary Economics*, November 1981, 8, 387-93.
- \_\_\_\_\_, "Real Exchange Rates under Alternative Nominal Exchange Rate Systems," *Journal of International Money and Finance*, August 1983, 2, 147-66.
- Svensson, Lars E.O., "Currency Prices, Terms of Trade, and Interest Rates: A General Equilibrium Asset-Pricing Cash-in-Advance Approach," *Journal of International Economics*, 1985 forthcoming.
- World Bank, *World Development Report 1981*, New York: Oxford University Press, 1982.

# Optimal Wage Indexation, Foreign Exchange Intervention, and Monetary Policy

By JOSHUA AIZENMAN AND JACOB A. FRENKEL\*

This paper deals with the design of optimal monetary policy and with the interaction between the optimal degrees of wage indexation and foreign exchange intervention. Recent studies of wage indexation in the closed economy have established that the optimal degree of wage indexation depends on the characteristics of the stochastic disturbances that affect the economy. In many of these studies, specifically in those that have adopted the analytical framework originated by Jo Anna Gray (1976), labor markets are characterized by the existence of nominal contracts that result in some stickiness of nominal wages. In these studies, indexation is intended to reduce the undesirable consequences of the stickiness of wages. Subsequent analyses of the optimal degree of wage indexation examined the implications of alternative assumptions about the determinants of employment in disequilibrium situations, as well as the rationale for the existence

of nominal contracts that yield sticky wages (see, for example, Robert Barro, 1977; Stanley Fischer, 1977a, b; Gray, 1978; Alex Cukierman, 1980; Edi Karni, 1983).

The analysis of optimal foreign exchange intervention, on the other hand, focused initially on the choice between a completely fixed and a completely flexible exchange rate system. Subsequent examinations of the same question have shifted the focus from the problem of choice between the two extreme exchange rate regimes to the problem of the optimal degree of exchange rate flexibility. Thus, the focus has shifted towards finding the *optimal mix* of the fixed and the flexible exchange rate regimes. Consequently, that analysis has attempted to determine the optimal degree of exchange rate management (see our 1982 article and the references therein).

More recently it has been recognized that the optimal degree of wage indexation depends on the prevailing exchange rate regime. Thus, Robert Flood and Nancy Marion (1982) showed that a small open economy with fixed exchange rates should adopt a policy of complete wage indexation whereas an economy with flexible exchange rates should adopt a policy of partial wage indexation. This analysis was extended by Aizenman (1985) who showed that, under flexible exchange rates, the optimal degree of wage indexation rises with the degree of openness of the economy as measured by the relative size of the traded goods sector. On the other hand, some authors have recognized that the choice between fixed and flexible exchange rate regimes depends on labor market conventions (for example, see Jagdeep Bhandari, 1982). Specifically, it has been argued that the degree of wage indexation determines the relative efficiency of macroeconomic policies under alternative exchange rate regimes and, therefore, the choice between the two re-

\*Graduate School of Business and the Department of Economics, respectively, University of Chicago, Chicago, IL 60637, and the National Bureau of Economic Research. A previous version of this paper was presented under the title "Wage Indexation and the Optimal Exchange Rate Regime," at the 1983 NBER Summer Institute, Cambridge, MA. We acknowledge helpful comments by G. Calvo, P. de Grauwe, S. Fischer, R. Flood, E. Helpman, R. Hodrick, C. Kahn, K. Kimbrough, P. Kouri, L. Leiderman, M. Mussa, M. Obstfeld, A. Razin, L. Weiss, Y. Weiss, and participants in seminars held at the NBER Summer Institute, Hebrew University, Tel-Aviv University, Columbia University, University of Chicago, Duke University, Northwestern University, Ohio State University, Johns Hopkins University, the University of Pennsylvania, the University of Rochester, University of California-Los Angeles, Vanderbilt University, the Board of Governors of the Federal Reserve System, and the International Monetary Fund. The research reported here is part of the NBER's research program in International Studies and Economic Fluctuations. Any opinions are our own and not those of the NBER.

gimes should depend on whether wages are indexed or not (for example, see Jeffrey Sachs, 1980, and Richard Marston, 1982a).

Common to these studies is the characteristic that the economy is either searching for the optimal degree of wage indexation under the assumption that the exchange rate regime (being fixed or flexible) is exogenously given, or that it is choosing between fixed and flexible exchange rate regimes under the assumption that the degree of wage indexation is exogenously given. The point of departure of this paper is the notion that the optimal degrees of wage indexation and exchange rate intervention are interrelated and are mutually and simultaneously determined. Therefore, in our analytical framework the choice of the optimal degrees of wage indexation and exchange rate intervention emerges as the outcome of a joint-optimization problem. This joint-optimization outcome is shown to be a component of the solution to the broader problem of the design of optimal monetary policy.

The interdependence among monetary policy, foreign exchange intervention, and labor market conditions, as characterized by the degree of wage indexation, has been clearly recognized by policymakers and has been viewed as an important constraint on the conduct of policy particularly in highly inflationary countries. And yet, except for few exceptions like Stephen Turnovsky (1983a), the question of the formal interaction between the optimal degrees of indexation and foreign exchange intervention, especially within the context of the design of optimal monetary policy has not received attention in the theoretical literature. This question is addressed in the subsequent sections.

Section I describes the building blocks of the model, including the determination of output and employment, the specification of wage contracts, and the determination of prices and exchange rates. One of the key characteristics of the model is the menu of the stochastic shocks. It is assumed that the economy is subject to stochastic shocks to productivity, to foreign prices, to purchasing power parities, to the rate of interest, and to the money supply. Much of the analysis de-

pends, therefore, on the relative magnitudes of these shocks, as well as on the information set that individuals are assumed to possess.

Our analysis assumes that, due to cost of negotiations, nominal wages are precontracted and real wages adjust according to a simple indexation formula that links the change in wages to the observed change in the price level. The level of employment in turn is assumed to be determined by firms according to their demand for labor. This specification of labor market conventions may result in discrepancies between the *realized* levels of real wages and employment and the *equilibrium* levels obtained when labor markets clear continuously without friction. The goal of policies is to minimize the welfare loss associated with such discrepancies.

Section II contains an analysis of the objective function which is given a formal justification in the Appendix. In Section III, we specify the optimal money supply process and derive the optimal set of policy rules that should govern the conduct of monetary policy. These policy rules determine the optimal response of monetary policy to changes in exchange rates, interest rates, and foreign prices. One of the key results is that the adoption of the optimal set of policy rules results in the complete elimination of the welfare cost. Thus, optimal policies nullify the distortions arising from the simple indexation rule and from the existence of nominal contracts. Since optimal policies succeed in the elimination of the distortions, critical issues concerning the nature of contracts and the implications of specific assumptions about disequilibrium positions become inconsequential. We then proceed to examine the interdependence between the design of optimal monetary policy rules and the optimal degree of wage indexation. This section concludes with the proposition that the number of independent indicators that govern a policy aiming at the elimination of a distortion, must equal the number of independent sources of information that influence the determination of the undistorted equilibrium. Thus, it is shown that with a sufficient number of indicators for monetary policy, there may be no need to introduce wage indexa-

tion. By the same token it is also shown that an economy that is not able to choose freely an exchange rate regime can still eliminate the welfare loss by supplementing the (constrained) monetary policy with an optimal rule for wage indexation.

Section IV examines the implications of the optimal policies on the means and the variances of money and output. In Section V, we apply our analytical framework to second-best situations in which policy cannot be used optimally. In this context we determine the optimal indexation coefficient for an economy that is constrained to follow a given exchange rate regime, and determine the optimal degree of exchange rate intervention for an economy that is constrained to follow a given wage indexation rule. For both of these cases we show the dependence of the (constrained) optimal policies on the details of the stochastic disturbances that affect the economy, and we compute the values of the loss function that result from the adoption of various policies.

### I. The Model

The model that we use has several building blocks. These include the specification of output and employment, the specification of the wage rule and the determination of prices and exchange rates. In this section we outline the structure of the model.

#### A. Output and Employment

Let the production function be

$$(1) \log Y_t = \log B + \beta \log L_t + \mu_t, \quad 0 \leq \beta \leq 1,$$

where  $Y_t$ ,  $L_t$ , and  $\mu_t$  denote, respectively, the level of output, the input of labor, and a productivity shock, at time  $t$ . The productivity shock  $\mu_t$  is assumed to be distributed normally with a zero mean and a known variance  $\sigma_\mu^2$ . Within each period the realized value of the productivity shock is not known, and the expectations concerning the realized value of  $\mu_t$  are formed on the basis of the information that is available during the period. Throughout the analysis we assume that at each point in time *all* prices and rates

of interest are known. The *conditional* expectation of  $\mu_t$ , as based on the information available at period  $t$ , is denoted by  $E_t(\mu_t)$ .

Producers are assumed to maximize the expected value of profits subject to the available information. Thus, in their demand for labor, producers are assumed to equate the real wage to the expected marginal product of labor. Expressed logarithmically, this equality implies that

$$(2) \log(W/P)_t = \log \beta B - (1 - \beta) \log L_t + E_t(\mu_t),$$

where  $W$  and  $P$  denote the nominal wage and the price level, respectively. From equation (2), the demand for labor is

$$(3) \log L_t^d = [-\log(W/P)_t + \log \beta B + E_t(\mu_t)] / (1 - \beta),$$

where  $L_t^d$  designates the demand for labor.<sup>1</sup> In order to simplify notation, we suppress from here on the subscript  $t$ . Thus, unless stated otherwise, the conditional expectation of the productivity shock  $E_t(\mu_t)$  will be denoted  $E(\mu)$ , and will also be referred to as the perceived productivity shock. Assuming that employment is determined by the demand for labor, we substitute equation (3) into (1) and obtain the level of output that

<sup>1</sup>Formally, the firm facing a given real wage is assumed to demand labor so as to maximize the expected value of profits conditional on the available information. Thus

$$\max_{L_t} E_t \{ BL_t^\beta e^{\mu_t} - (W/P)_t L_t \}.$$

The resulting demand for labor (expressed logarithmically) is

$$\log L_t^d = [\log \beta B - \log(W/P)_t + \log E_t(e^{\mu_t})] / (1 - \beta),$$

and using the approximation  $\log E_t(e^{\mu_t}) \approx E_t(\mu_t)$  we obtain equation (3). The same approximation, that is valid for small values of the variance and the realization of the stochastic shock, is also used in the derivation of the expected value of the marginal product of labor in equation (2).

corresponds to the employment of labor:

$$(4) \quad \log Y = \log B + \beta\sigma [\log P - \log W \\ + \log \beta B + E(\mu)] + \mu,$$

where  $\sigma \equiv 1/(1-\beta)$ . Equation (4) specifies the *stochastic* supply of output that is obtained when the value of the productivity shock is  $\mu$ . The corresponding *deterministic* level of output is

$$(4') \quad \log Y_0 = \log B \\ + \beta\sigma (\log P_0 - \log W_0 + \log \beta B),$$

where  $P_0$  and  $W_0$  denote the market-clearing price level and nominal wage that are obtained in the absence of stochastic shocks.

For the subsequent analysis it is useful to denote by lowercase letters the percentage discrepancy of a variable from the value that it obtains in the absence of shocks. Thus,  $x \equiv \log X - \log X_0$ . Accordingly, from equations (4) and (4') we obtain

$$(5) \quad y = \beta\sigma [p - w + E(\mu)] + \mu.$$

Equation (5) shows that the percentage deviation of output from its deterministic level depends on the percentage deviation of the real wage from its deterministic value, on the perceived productivity shock,  $E(\mu)$ , as well as on the realized productivity shock,  $\mu$ . Our subsequent analysis specifies the determinants of  $w$ ,  $p$ , and  $E(\mu)$ .

### B. The Wage Rule

It is assumed that due to costs of negotiations, nominal wages are set according to the following simple, time-invariant, indexation rule:

$$(6) \quad \log W_t = \log W_0 + b(\log P_t - \log P_0).$$

Equation (6) specifies the wage at period  $t$  as a function of  $W_0$ , the equilibrium wage that would have prevailed if shocks were zero, and the percentage deviation of the price

from its nonstochastic value.<sup>2</sup> In equation (6),  $b$  designates an indexation parameter. When  $b=1$ , wages are fully indexed to the rate of inflation and the real wage is rigid. When  $b=0$ , nominal wages are rigid. From equation (6) it follows that  $w = bp$ . Substituting  $bp$  for  $w$  in equation (5) yields

$$(7) \quad y = \beta\sigma [(1-b)p + E(\mu)] + \mu.$$

Equation (7), which may be viewed as an aggregate supply function, expresses the supply of output as a function of the price  $p$ , as well as the perceived and realized productivity shocks. The dependence of the supply on the price depends in turn on the coefficient of indexation  $b$ ; a higher indexation coefficient results in a weaker dependence of output on the price.

### C. The Price Level and the Exchange Rate

The domestic price level is assumed to be linked to the foreign price through purchasing power parity. Let the foreign price be

$$(8) \quad \log P'_t = \log \bar{P} + \chi_t,$$

where a prime denotes a foreign variable and a bar over a variable denotes the value of its fixed component. In equation (8),  $\chi_t$  denotes the stochastic component of the foreign price which is assumed to be distributed normally with zero mean and a fixed known variance. The domestic price is linked to the foreign price according to

$$(9) \quad \log P_t = \log S_t + \log P'_t,$$

where  $S_t$  denotes the exchange rate (the price of foreign currency in terms of domestic

<sup>2</sup>It is assumed that the initial nominal wage is set at the level  $W_0$ . This assumption is justified in the Appendix. Our specification of the indexation rule corresponds to a wage rule that is widely used in practice, and its main virtue is simplicity. Much of our subsequent analysis aims to demonstrate that with proper monetary policy, which is governed by time-invariant policy rules, this simplicity need not yield suboptimal outcomes.

currency). Using (8) for  $\log P_t'$  yields

$$(10) \quad \log P_t = \log S_t + \log \bar{P}' + \chi_t.$$

In principle, the random component of  $P_t$  may also include stochastic deviations from the purchasing power parity relation of equation (9). When all shocks are zero, the domestic price is

$$(10') \quad \log P_0 = \log S_0 + \log \bar{P}',$$

and subtracting (10') from (10) yields

$$(11) \quad p = s + \chi.$$

where, as before, we suppress the time subscripts.

The formulation in equation (11) links the domestic price to the exchange rate and the stochastic shock  $\chi$ . In order to determine the level of prices we need to incorporate monetary considerations. The equilibrium price level and exchange rates can be derived from the conditions of money market equilibrium. Let the demand for money be

$$(12) \quad \log M_t^d = \log K + \log P_t + \log Y_t - \alpha i_t,$$

where  $M$  denotes nominal balances and  $i$  denotes the nominal rate of interest. The nominal rate of interest in turn is linked to the foreign rate of interest,  $i'$ . Arbitrage by investors, who are assumed to be risk neutral, assures that uncovered interest parity holds:<sup>3</sup>

$$(13) \quad i_t = i'_t + E_t(\log S_{t+1} - \log S_t),$$

where  $E_t \log S_{t+1}$  denotes the expected exchange rate for period  $t+1$  based on the information available at period  $t$ . The foreign rate of interest is also subject to a random shock,  $\rho$ , which is distributed normally with zero mean and a fixed known

variance. Thus,

$$(14) \quad i'_t = \bar{i}' + \rho_t.$$

In specifying the money supply process, we assume that the monetary authority takes account of the relevant *information* conveyed by a specific set of independent indicators. Since at each point in time prices and interest rates are known, we assume that the supply of nominal balances adjusts in response to the three independent indicators  $s_t$ ,  $\rho_t$ , and  $\chi_t$ , according to

$$(15) \quad \log M_t^s = \log \bar{M} + \delta_t - \gamma s_t - \tau \rho_t - \xi \chi_t$$

where  $\delta$  (which is assumed to be distributed normally with zero mean and a fixed known variance) denotes a random shock to the money supply process. In equation (15),  $\gamma$  denotes the elasticity of the money supply with respect to  $s$ —the deviation of the exchange rate from its deterministic value,  $\tau$  denotes the elasticity of the money supply with respect to  $\rho$ —the stochastic shock to the foreign rate of interest, and  $\xi$  denotes the elasticity of the money supply with respect to  $\chi$ —the stochastic shock to foreign prices. In the subsequent analysis of the money supply rule, we justify the choice of this set of indicators and determine the optimal values of the time-invariant coefficients  $\gamma$ ,  $\tau$ , and  $\xi$ .

Equilibrium in the money market requires that

$$(16) \quad \log K + \log P_t + \log Y_t - \alpha i_t = \log \bar{M} + \delta_t - \gamma s_t - \tau \rho_t - \xi \chi_t,$$

and, when all shocks are zero, money market equilibrium yields<sup>4</sup>

<sup>4</sup> It is relevant to note that from equations (13)–(14),

$$i - \bar{i}' = \rho + E_t \log S_{t+1} - \log S_t,$$

and the specification of the stochastic shocks implies that  $E_t \log S_{t+1} = \log S_0$ . The implicit assumption underlying this formulation is that  $E_t \log S_{t+1}$  is not influenced by the observed price. Our assumption about the absence of trend allows us to focus on the properties of the stationary equilibrium for which the current

<sup>3</sup> More precisely, when prices are stochastic, uncovered interest parity holds only as an approximation due to Jensen's inequality. This approximation is valid for small values of the variance of the stochastic shock to prices; see Frenkel and Assaf Razin (1980).



$$(16') \quad \log K + \log P_0 + \log Y_0 - \alpha \bar{i}' = \log \bar{M}.$$

Subtracting (16') from (16) and omitting the time subscript yields

$$(17) \quad p + y - \alpha(i - \bar{i}') = \delta - \gamma s - \tau p - \xi \chi.$$

Substituting equations (7) and (11) for  $y$  and  $p$  and using the fact that the domestic rate of interest (from equations (13)–(14)) is  $i' + \rho - s$ , we obtain

$$(18) \quad \lambda(s + \chi) + \beta \sigma E(\mu) + \mu - \alpha(\rho - s) \\ = \delta - \gamma s - \tau p - \xi \chi,$$

where  $\lambda \equiv [1 + \beta \sigma(1 - b)]$ .

In equation (18),  $\lambda$  denotes the elasticity of nominal income (and thereby of the reduced-form demand for money) with respect to prices. As may be seen, the magnitude of  $\lambda$  depends on the size of the indexation coefficient  $b$ . When wage indexation is complete (i.e., when  $b = 1$ ), price changes do not alter output and  $\lambda = 1$ . When  $b$  is less than unity, a rise in the price alters real wages by  $(1 - b)$  and, therefore, it also affects money demand through changing real output by  $\beta \sigma(1 - b)$ . From equation (18) it follows that the equilibrium percentage change in the exchange rate is

$$(19) \quad s = \frac{(\alpha - \tau)\rho - (\mu - \delta) - \beta \sigma E(\mu) - (\lambda + \xi)\chi}{\lambda + \alpha + \gamma}.$$

As is evident from equation (19), when  $\gamma = 0$  the exchange rate is fully flexible, and when  $\gamma = \infty$ ,  $s = 0$  and the exchange rate is fixed. Between these two extremes there is a wide range of intermediate exchange rate regimes.

Recalling that  $p = s + \chi$ , and using equation (19) we can express the price as

$$(20) \quad p = \chi + \frac{(\alpha - \tau)\rho - (\mu - \delta) - \beta \sigma E(\mu) - (\lambda + \xi)\chi}{\lambda + \alpha + \gamma}.$$

As may be seen, the price depends on the realization of the stochastic shocks, on the perceived value of the real shock  $E(\mu)$ , on the coefficient of wage indexation  $b$ , on the coefficient of foreign exchange intervention  $\gamma$ , and on the other coefficients which govern monetary policy.

The value of  $E(\mu)$  that is consistent with the information structure and with the requirement of rational expectations reflects the information set that is available to decision makers. We assume that at each point in time individuals observe the current values of the price  $p$ , the exchange rate  $s$ , and the rates of interest  $i$  and  $i'$ , but they cannot observe *directly* the stochastic shocks. Since our analysis does not deal with issues arising from asymmetric information, we also assume that individuals know the policy rules. The available information set can be used by individuals in order to infer the values of some of the shocks. For example, the observed values of  $p$  and  $s$  imply the value of  $\chi$  (from equation (11)), and the observed value of  $i'$  implies the value of  $\rho$  (from equation (14)). While individuals do not possess knowledge about the values of the real productivity shock  $\mu$ , and the money supply shock  $\delta$ , their knowledge of the values of  $p$ ,  $\chi$ , and  $\rho$  along with their knowledge of the coefficient of wage indexation  $b$ , and of the coefficients of the money supply function  $\gamma$ ,  $\tau$ , and  $\xi$ , implies, from equation (20), a value for  $(\mu - \delta)$ . The value of  $(\mu - \delta)$  may be viewed as the informational content of the price  $p$ . The assumption of rational expectations implies that the optimal forecast of  $\mu$  reflects an efficient use of this information. Thus, the value of  $E(\mu)$  may be computed from a regression of  $\mu$  on  $(\mu - \delta)$ . The ordinary least squares estimate of the real shock that is obtained through this procedure is  $E(\mu) = \psi(\mu - \delta)$ , where  $\psi \equiv \text{cov}(\mu, \mu - \delta) / \sigma_{(\mu - \delta)}^2$ , and where  $\sigma_{(\mu - \delta)}^2$  de-

values of the stochastic shocks do not affect the expectations about future values of the variables. The specification of equation (16') also embodies the assumption that the equilibrium is unique. The choice of the unique equilibrium is consistent with the criterion suggested by Bennett McCallum (1983). On the issue of uniqueness, see Guillermo Calvo (1979), and Turnovsky (1983b).

notes the variance of  $(\mu - \delta)$ .<sup>5</sup> When the shocks are independent of each other the regression coefficient becomes  $\psi = \sigma_\mu^2 / (\sigma_\mu^2 + \sigma_\delta^2)$ , where the variance of a variable,  $x$ , is denoted by  $\sigma_x^2$  (to be distinguished from the production parameter  $\sigma \equiv 1/(1 - \beta)$ ).

Finally, substituting the estimates of the real shock  $E(\mu)$  into equation (20), we obtain

$$(21) \quad p = \frac{(\alpha - \tau)\rho + (\alpha + \gamma - \xi)\chi - (1 + \beta\sigma\psi)(\mu - \delta)}{\lambda + \alpha + \gamma}$$

This solution for  $p$  can be substituted into equation (7) to yield an expression for the aggregate supply as a function of the stochastic structure of the shocks, the coefficient of wage indexation  $b$  (that is embodied in the value of  $\lambda$ ), and the various coefficients which govern policy. In order to determine the optimal values of these coefficients we turn next to an analysis of the objective function.

## II. The Objective Function

The foregoing analysis determined the level of output,  $y$  (or more precisely the percentage deviation of output from the level that would have prevailed in the absence of shocks) under the assumption that employment is determined exclusively by the demand for labor (equation (3)). The resultant disequilibrium in the labor market induces welfare cost. We assume that the policy goal is to minimize this welfare cost by choosing the optimal values of the coefficient of indexation and of the coefficients governing monetary policy.

In order to compute the welfare cost, we compute the level of employment  $\tilde{L}$  that

would have prevailed under conditions of full clearance of labor markets. We then compare  $\tilde{L}$  with the actual level of employment  $L$ , and compute the welfare cost that is associated with a discrepancy between  $\tilde{L}$  and  $L$ .<sup>6</sup>

Let the supply of labor be

$$(22) \quad \log L_t^s = \log A + \varepsilon \log (W/P)_t,$$

where  $\varepsilon$  denotes the elasticity of labor supply. Equating the supply of labor, equation (22), with the demand for labor, equation (3), yields the *equilibrium* level of employment,  $\log \tilde{L}$ , where

$$(23) \quad \log \tilde{L} = \log A + \varepsilon [(\sigma(E(\mu) + \log \beta B) - \log A) / (\sigma + \varepsilon)],$$

and subtracting from (23) the equilibrium level of employment that would have prevailed in the absence of shocks, we obtain

$$(24) \quad \tilde{l} = \varepsilon \sigma E(\mu) / (\sigma + \varepsilon).$$

Actual employment, however, may not adjust to clear labor markets; rather, it is governed by the assumptions that labor is demand determined and wages are determined by the indexation rule. Subtracting from the actual supply of output (equation (1)) the supply that would have obtained in the absence of shocks, yields

$$(25) \quad y = \beta l + \mu,$$

and thus, employment (or more precisely the percentage deviation of employment from the level that would have prevailed in the absence of shocks) is

$$(26) \quad l = (y - \mu) / \beta.$$

<sup>5</sup>This procedure for determining  $E(\mu)$  may be viewed as a short cut to the more lengthy computation following the undetermined coefficients method. An analogous short cut is adopted in Matthew Canzoneri, Dale Henderson, and Kenneth Rogoff (1983) in the context of an analysis of the informational content of interest rates.

<sup>6</sup>A formal derivation of the loss function is presented in the Appendix in terms of utility maximization. In what follows we provide a somewhat less formal exposition in terms of consumers' and producers' surplus.

By using equation (5) for the value of  $y$ ,  $l$  can be written as

$$(26') \quad l = -\sigma[(w - p) - E(\mu)],$$

and, therefore, the discrepancy (in percentage terms) between equilibrium and actual employment is

$$(27) \quad l - \bar{l} = \sigma \left[ -(w - p) + \frac{\sigma}{\sigma + \varepsilon} E(\mu) \right].$$

In order to compute the welfare loss associated with this discrepancy, we need to multiply it by one-half of the difference between the demand and the supply prices at the actual employment level. This procedure amounts to a computation of the area of a triangle representing the welfare cost in terms of the loss of consumers' and producers' surplus. From the demand for and the supply of labor, these demand and supply prices (or more precisely the percentage changes thereof) are, respectively,

$$(28) \quad (w - p)^d = -(l - \bar{l})/\sigma + (\widetilde{w - p})$$

$$(29) \quad (w - p)^s = (l - \bar{l})/\varepsilon + (\widetilde{w - p}),$$

and the percentage welfare cost of suboptimal employment is therefore

$$(30) \quad \frac{\sigma^2}{2} \left( \frac{1}{\varepsilon} + \frac{1}{\sigma} \right) \left[ -(w - p) + \frac{\sigma E(\mu)}{\sigma + \varepsilon} \right]^2.$$

As is clear from equation (30), once we omit the irrelevant constants, minimizing the expected welfare loss on the basis of the information available at period  $t - 1$  amounts to minimizing the loss function  $H$ :

$$(31) \quad H = E \left[ \left\{ -(w - p) + \left( \sigma / (\sigma + \varepsilon) \right) E(\mu) \right\}^2 \middle| I_{t-1} \right],$$

where  $I_{t-1}$  denotes the information set available at period  $t - 1$ .<sup>7</sup>

<sup>7</sup>It is relevant to note that the formulation of the objective function in terms of a minimization of the welfare cost of the distortions in the labor market is

### III. Optimal Policies

In order to find the optimal values of the coefficient of indexation and the other feedback coefficients that govern policy, we substitute into equation (31) the indexation rule  $w = bp$ , the forecasting rule  $E(\mu) = \psi(\mu - \delta)$ , and the solution for  $p$  from equation (21), and obtain the following loss function:

$$(32) \quad H = E \left[ \left\{ \phi \theta + \left( \sigma / (\sigma + \varepsilon) \right) \psi(\mu - \delta) \right\}^2 \middle| I_{t-1} \right],$$

where

$$(33) \quad \begin{cases} \phi \equiv ((1 - b)/(\lambda + \alpha + \gamma)) \\ \theta \equiv (\alpha - \tau)\rho + (\alpha + \gamma - \xi)\chi \\ \quad - (1 + \beta\sigma\psi)(\mu - \delta). \end{cases}$$

In equation (32),  $\phi\theta$  denotes the (percentage) change in the real wage  $(1 - b)p$ .

In interpreting the loss function (32), it is useful to note that the term  $\psi(\mu - \delta)$  de-

equivalent to the more conventional (but somewhat less informative) formulation of minimizing the expected squared discrepancy of output,  $y$ , from the equilibrium level,  $\bar{y}$ , obtained with full market clearing (see Aizenman, 1983). This equivalence becomes evident by noting that since  $(y - \bar{y}) = \beta(l - \bar{l})$ ,  $E(y - \bar{y})^2 = \beta^2 E(l - \bar{l})^2$ . In order to obtain the welfare loss in units of output we need to multiply equation (30) by the equilibrium real wage bill,  $(W/P)\bar{L}$ . The resulting quantity is the same as equation (A8) in the Appendix. Our focus on the labor market in the computation of the welfare cost presumes that other markets are undistorted. An explicit incorporation of this assumption would require that monetary policies at home and abroad generate the optimal rate of inflation. With this interpretation, our formulation of the stochastic shock to the money supply would be viewed as a random deviation from the deterministic trend reflecting the optimal rate of inflation. Equation (31) presumes that the authorities aim to minimize the expectations of the welfare loss on the basis of the information set available at period  $t - 1$ . The alternative specification which minimizes the expected welfare loss on the basis of the currently available information yields policy coefficients that are not time-invariant. Since, as will be seen below, the optimal time-invariant coefficients eliminate the welfare loss, the choice between the two procedures might reflect the excess costs associated with state-dependent rule.

notes the private sectors' optimal forecast of the real shock,  $\mu$ , and its product with  $\sigma/(\sigma + \varepsilon)$  measures therefore the *equilibrium* change in real wages that would occur under an optimal use of information. On the other hand, the *actual* change in real wages that results from the adoption of specific policy rules is  $-\phi\theta$ . The expected squared discrepancy between the two magnitudes, that is, the variance of the error in the determination of actual real wages, entails welfare loss which is measured by the loss function  $H$ . By inspecting the value of  $\theta$  in equation (33), it is clear that in order to minimize the loss function (32), we need to set

$$(34) \quad \tau^* = \alpha;$$

$$(35) \quad \xi^* = \alpha + \gamma,$$

where  $\tau^*$  and  $\xi^*$  designate the optimal values of  $\tau$  and  $\xi$ . Assuming that the values of  $\tau$  and  $\xi$  are set according to equations (34) and (35), the loss function (32) becomes

$$(32') \quad E \left\{ \left[ -\phi(1 + \beta\sigma\psi) + (\sigma/(\sigma + \varepsilon))\psi \right] (\mu - \delta) \right\}^2 | I_{t-1} \right\},$$

and it is evident that the value of  $\phi$  which equates (32') to zero is

$$(36) \quad \phi^* = \sigma\psi / [(\sigma + \varepsilon)(1 + \beta\sigma\psi)].$$

Finally, by equating the value of  $\phi^*$  with its definition in equation (33), we solve for the optimal value of  $\gamma$ :

$$(37) \quad \gamma^* = (1 - b)((\sigma + \varepsilon)/\sigma\psi)(1 + \beta\sigma\psi) - \alpha - \lambda,$$

and substituting  $1 + \beta\sigma(1 - b)$  for the value of  $\lambda$  yields

$$(37') \quad \gamma^* = (1 - b)[(\sigma + \varepsilon(1 + \beta\sigma\psi))/\sigma\psi] - (1 + \alpha).$$

Equations (34), (35), and (37') provide three restrictions on the values of the four policy coefficients  $\tau$ ,  $\xi$ ,  $\gamma$ , and  $b$ . As is

evident, this set of restrictions contains one degree of freedom. Since, however, the structure of the model implies that these restrictions are recursive, it follows that of the four policy coefficients,  $\tau$  is indispensable. The degree of freedom permits setting an arbitrary value to one of the coefficients in the triplet  $(b, \xi, \gamma)$ , while setting the other two at their optimal values. For example, if the indexation coefficient  $b$  is given exogenously, the restrictions in equations (34), (35), and (37') imply the optimal values of  $\tau$ ,  $\xi$ , and  $\gamma$ . This provides the rationale for the specification of the money supply process in equation (15). Adopting this optimal set of policy rules for the money supply process results in the *elimination* of the welfare loss.

Equation (37') also suggests that the optimal value of  $\gamma$  depends on the structural parameters of the economy (including the semi-elasticity of the demand for money ( $\alpha$ ), the elasticity of output with respect to labor input ( $\beta$ ), and the elasticity of the supply of labor ( $\varepsilon$ )); on the stochastic structure of the real and the monetary shocks (that govern the value of  $\psi$ ) and on the indexation coefficient  $b$ . Thus, for example, the higher the elasticity of the supply of labor, the larger becomes the optimal value of  $\gamma$ , that is, the larger becomes the desirability of greater fixity of exchange rates.

As is evident from equation (37'), around the optimum, there is a negative correlation between the value of  $\gamma^*$  and the (exogenously given) degree of wage indexation. Thus, an economy with a higher degree of wage indexation will find it optimal to increase the flexibility of exchange rates (reduce  $\gamma^*$ ). As the coefficient of wage indexation approaches unity, the degree of real wage rigidity increases, and the optimal value of  $\gamma^*$  approaches  $-(1 + \alpha)$ .<sup>8</sup> Furthermore,

<sup>8</sup>At the extreme, with full indexation, the optimal value of  $\gamma$  is undetermined. This may be verified by reference to the loss function in equations (32)-(33), where it is seen that when  $b = 1$ , the value of  $\phi$  is zero and, as a result, the value of the loss function is independent of  $\gamma$ . Intuitively, full indexation introduces real wage rigidity. Consequently, changes in the price level that can be brought about through changes in the exchange rate and that are influenced by the exchange rate

since from equation (35) the value of  $\xi^*$  depends linearly on  $\gamma^*$ , it also follows that a higher degree of wage indexation lowers the optimal degree to which monetary policy responds to  $\chi$  (the shocks to foreign prices).

The forgoing analysis also demonstrates that as long as the money supply responds optimally to  $s$ ,  $\rho$ , and  $\chi$ , which in the present case are the relevant sources of independent information that can be used to yield the market-clearing real wage, there is no need to introduce wage indexation. Thus, it was shown that when the degree of freedom provided by equations (34), (35), and (37) is used up by setting the indexation coefficient at an exogenously given level, the welfare loss can be eliminated by a proper choice of  $\gamma$ ,  $\tau$ , and  $\xi$ . If, on the other hand, the value of  $\gamma$  were given exogenously, then the welfare loss could still be eliminated by supplementing the optimal values of  $\tau$  and  $\xi$  in the money supply process with an optimal rule of wage indexation. From equation (37') the optimal value of  $b$  for an exogenously given value of  $\gamma$  is

$$(38) \quad b^* = 1 - (1 + \alpha + \gamma) / \left[ (\sigma_s^2 / \sigma_\mu^2) (1 + (\varepsilon / \sigma)) + 1 + \varepsilon \right].$$

The dependence of the value of  $b^*$  on the magnitudes of the key parameters is qualitatively similar to the dependence of  $\gamma^*$  on these parameters. Thus

$$\frac{\partial b^*}{\partial \psi} < 0, \quad \frac{\partial b^*}{\partial \varepsilon} > 0, \quad \frac{\partial b^*}{\partial \gamma} < 0, \quad \frac{\partial b^*}{\partial \beta} < 0.$$

Accordingly, a rise in the relative variance of the real shock, a rise in the elasticity of output with respect to labor input, and a rise in the degree of fixity of exchange rates result in a lower optimal value of the indexation coefficient, whereas a rise in the elastic-

ity of labor supply raises the optimal degree of wage indexation. It is relevant to note that by setting  $\gamma = 0$ , the optimal indexation coefficient becomes

$$(38') \quad b_c^* = 1 - (1 + \alpha) / \left[ (\sigma_s^2 / \sigma_\mu^2) (1 + (\varepsilon / \sigma)) + 1 + \varepsilon \right],$$

where  $b_c^*$  denotes the closed-economy value. This is indeed the optimal indexation coefficient that is derived in Aizenman's (1983) closed-economy model.<sup>9</sup>

The economic intuition underlying the redundancy of one of the coefficients in the triplet  $(\gamma, \xi, b)$  is implicit in the structure of the model. Since the rate of interest appears only in the demand for money, the only way of eliminating the impact of an interest rate shock on the loss function is by setting  $\tau^* = \alpha$  as in equation (34). No other policy rule can eliminate the impact of an interest rate shock. In contrast, the rest of the shocks manifest themselves through the price level and, together with the given nominal wage, they impact on the real wage which is the source of the welfare loss. Since from equation (21),  $\gamma$  and  $\xi$  influence the price level whereas the wage indexation coefficient influences both the price level and the nominal wage, they all alter the real wage directly. Given the nature of the shocks we need only three independent indicators.<sup>10</sup> Therefore, it is sufficient to

<sup>9</sup>The intuition underlying this result is that the optimal policy-response coefficients for the open economy ensure that the price level effects arising from the shocks  $\rho$  and  $\chi$  (that originate from the openness of the economy) are offset by setting  $\tau = \alpha$  and  $\alpha + \gamma = \xi$ . Thus, at the optimum, policy succeeds in creating an outcome that is equivalent to the one generated by  $\rho = \chi = 0$ . Since in the closed economy  $\rho = \chi = \gamma = 0$ , we only need to substitute  $\gamma = 0$  in equation (38) to obtain the closed-economy result. Alternatively,  $b_c^*$  can be obtained directly from the loss function (32)–(33) by noting that when the economy is closed,  $\rho = \chi = \gamma = 0$ ,  $\theta = -(1 + \beta\sigma\psi)(\mu - \delta)$ , and the value of  $b$  that eliminates the welfare loss is  $b_c^*$  as in (38').

<sup>10</sup>An analogous redundancy proposition is developed in Canzoneri, Henderson, and Rogoff in connection with the usage of the information contained in nominal interest rates. It is noteworthy that our objective function presumes that the only policy objective is the

regime will be inconsequential since, due to the rigidity of real wages, they will induce equiproportionate changes in nominal wages. Obviously, when real wages are completely rigid, monetary policy cannot eliminate labor market distortions arising from disequilibrium real wages.

use in addition to  $\tau$ , which is in this model an indispensable policy rule, any other pair from the triplet  $(\gamma, \xi, b)$ .

The examples analyzed above illustrated the substitutability between exchange rate flexibility and wage indexation under the assumptions either  $\gamma$  or  $b$  are set exogenously, and that  $\tau$  and  $\xi$ —the coefficients of response to interest rate shocks  $\rho$ , and to foreign price shocks  $\chi$ —are set optimally. Suppose now that the authorities do not adopt a policy response to  $\chi$ . Under these circumstances, again  $\tau^* = \alpha$  and, since  $\xi = 0$ , it follows from equation (35) that the optimal value of  $\gamma$  is  $-\alpha$ . Thus, when  $\xi = 0$  the solution for the optimal exchange rate regime is unique and, in contrast with the case described by equation (37'), the value of  $\gamma^*$  is independent of the deterministic and the stochastic structure of the economy.

The optimal value of the indexation coefficient corresponding to that case can be found from equation (38). Substituting  $\gamma = -\alpha$  yields

$$(38'') \quad b^* \Big|_{\substack{\tau^* = \alpha \\ \gamma^* = -\alpha}} \\ = 1 - 1 / \left[ \left( \sigma_\delta^2 / \sigma_\mu^2 (1 + (\varepsilon / \sigma)) + 1 + \varepsilon \right) \right].$$

A comparison of equation (38'') with (38') reveals that

$$(39) \quad b^* \Big|_{\substack{\tau^* = \alpha \\ \gamma^* = -\alpha}} > b_c^*.$$

That is, when in the open economy  $\tau$ ,  $\gamma$ , and  $b$  are set at their optimal values, the resultant wage indexation coefficient is larger than the corresponding closed-economy optimal indexation coefficient.<sup>11</sup>

elimination of distortions. If, in addition, the policy-maker wishes to reduce the variance of prices, then the redundant coefficient could be employed in the attainment of that target. In the present model, however, the assumption concerning the utility function does not provide an obvious rationale for reducing the variance of prices.

<sup>11</sup> This result reflects our specification of the nature of the shocks by which the openness of the economy does not increase the exposure to foreign real shocks. In

The incorporation of the various shocks as components of the policy rules governing the money supply process may serve to supplement Tinbergen's theorem concerning the relation between targets and instruments of economic policy. In our case the single "target" for economic policy is the elimination of a distortion to the real wage. This single target can be attained by means of the single instrument of monetary policy. Our analysis shows that the single instrument is capable of attaining the target only if it is triggered by a sufficient number of independent indicators. This number of independent indicators for the policy rules must equal the number of independent sources of information that influence the determination of the undistorted real wage. This perspective on the concept of optimal policy was illustrated in our model in terms of the characteristics of the money supply process. It does, however, have relevance for a wider range of policies including the characteristics of fiscal spending.

Finally, we have argued that the optimal policy could follow a sophisticated money supply rule that is triggered by a sufficient number of independent indicators. Alternatively, the optimal policy could follow a sophisticated wage indexation formula that is not limited to respond only to changes in the price level. Following the general principle, such an indexation formula will be optimal only if it responds to a sufficient number of signals and, as was argued before, the number of such independent indicators must equal the number of the independent sources of information that matter in determining the market-clearing real wage.<sup>12</sup> The choice among the alternatives of a sophisticated money supply rule, a sophisticated wage in-

principle the relative importance of real shocks may be higher for the open economy if, for example, it faces shocks to the price of imported raw materials. In that case the optimal indexation coefficient may be lower than  $b_c^*$ .

<sup>12</sup> For illustrations of the optimal design of sophisticated indexation formula in the context of a closed economy, see Fischer (1977a), Karni, and our 1985 paper.

dexation formula, or any other sophisticated set of policies is likely to be governed by the relative costs and complexities associated with each alternative. Such costs may reflect the difficulties of prompt implementations of alternative policy responses. The choice among alternative policies is also likely to be influenced by external constraints (like the rules of the IMF on foreign exchange intervention) and domestic institutional constraints (like the relative strength of the monetary authority and labor unions). Therefore, the actual choice of policy is likely to differ across different countries. The possibility that the optimal choice may differ across countries may be relevant for the design of IMF programs. Furthermore, from the global perspective, such possible international differences of optimal intervention rules need to be reconciled with each other in order to ensure consistency of the international monetary system.

#### IV. The Optimal Levels and Variability of Money and Output

In this section we assume that the optimal policies have been adopted and we examine the implications of these optimal policies on the means and the variances of the money supply and output.

##### A. The Optimal Money Supply

The money supply function was specified in equation (15) ( $m = \delta - \gamma s - \tau p - \xi \chi$ ). Substituting the optimal values of  $\tau$  and  $\xi$  from equations (34)–(35), and recalling that  $p = s + \chi$ , yields

$$(40) \quad m = \delta - \gamma p - \alpha(\rho + \chi).$$

Substituting (37') for the optimal value of  $\gamma$ , collecting terms and recalling that  $\rho + \chi - p = i - i'$  yields the optimal money supply:

$$(40') \quad m^* = \delta - [(1-b)((\sigma + \epsilon(1 + \beta\sigma\psi)) / \sigma\psi) - 1] p - \alpha(i - i').$$

Equation (40') which may be interpreted as a reduced-form optimal money supply,

expresses the dependence of  $m^*$  on the price and on the rate of interest. As may be seen, a rise in the rate of interest triggers a reduction in the money supply. The optimal reduction in the money supply aims to restore money market equilibrium and thereby to neutralize the effect of the change in the rate of interest on the price and, through it, on the real wage. Therefore, the (semi) elasticity of  $m^*$  with respect to  $i$  is  $-\alpha$ , and changes in  $m^*$  exactly match and offset changes in the demand for money.<sup>13</sup> The response of the optimal money supply to changes in  $p$  is more involved since it depends on the stochastic structure and on the coefficient of indexation.

In order to obtain further understanding of the characteristics of the optimal money supply, it is convenient to express  $m^*$  as a function of the stochastic shocks. For this purpose we note from equation (21) that with optimal policies the optimal price is<sup>14</sup>

$$(21') \quad p^* = \frac{-\sigma\psi}{(1-b)(\sigma + \epsilon)}(\mu - \delta).$$

Substituting equations (37') and (21') for the optimal values of  $\gamma$  and  $p$  into equation (40) and collecting terms yields

$$(40'') \quad m^* = -g\psi\delta + (1 + g\psi)\mu - \alpha(\rho + \chi),$$

$$\text{where} \quad g = \frac{\sigma}{\sigma + \epsilon} \left[ \beta\epsilon - \frac{1 + \alpha}{1 - b} \right].$$

<sup>13</sup> This property of the optimal money supply reflects the assumption that the rate of interest does not affect the real equilibrium of the economy. In a more elaborate framework the rate of interest may affect the real equilibrium through altering the supply of labor or through its impact on relative commodity prices.

<sup>14</sup> It may be shown that at the optimum the variance of the price is

$$\sigma_p^2 = (\sigma/(\sigma + \epsilon))^2 \psi \sigma_\mu^2 / (1 - b)^2.$$

In general,  $\sigma_p^2 = \sigma_s^2 + \sigma_\chi^2 + 2\text{cov}(s, \chi)$  and, since at the optimum under our assumptions  $\text{cov}(s, \chi) = -\sigma_\chi^2$ , it follows that  $\sigma_s^2 = \sigma_p^2 + \sigma_\chi^2$ , i.e., the variance of the exchange rate exceeds the variance of the price level.

The economic interpretation of (40'') is facilitated by substituting the perceived value of the real shock,  $E(\mu)$ , for  $\psi(\mu - \delta)$  and by rewriting (40'') as

$$m^* - \delta = (\mu - \delta) + gE(\mu) - \alpha(\rho + \chi),$$

where  $m^* - \delta$  denotes the optimal money supply net of the random component  $\delta$ . Thus,  $m^* - \delta$  is the part of the money supply that is attributed to the optimal policy rules. As may be seen, the parameter  $g$  is the elasticity of the optimal money supply with respect to the perceived value of the real shock. The sign of this elasticity depends on the coefficient of indexation and on the values of the structural parameters.

Using equation (40''), the variance of the optimal money supply can be written as

$$(41) \quad \sigma_{m^*}^2 = [1 + g(2 + g)\psi] \sigma_\mu^2 + \alpha^2 \sigma_{\rho+\chi}^2.$$

From equation (41) it is evident that a rise in the variance of  $\rho$ ,  $\chi$ , and  $\mu$  raises the variance of the optimal money supply while a rise in the variance of  $\delta$  exerts an ambiguous effect.

### B. Optimal Output

The level of output corresponding to the optimal policies is  $y^*$  which equals the level of output obtained with full market clearing. This level can be found from equation (24) and (25) or, alternatively, it can be found by substituting the optimal price from equation (21') into the aggregate supply in equation (7). Thus,

$$(42) \quad y^* = \frac{\beta\epsilon\sigma}{\sigma + \epsilon} \psi(\mu - \delta) + \mu.$$

From equation (42) it follows that at the optimum the variance of output is

$$(43) \quad \sigma_{y^*}^2 = \left[ 1 + \frac{\beta\epsilon\sigma}{\sigma + \epsilon} \psi \left( 2 + \frac{\beta\epsilon\sigma}{\sigma + \epsilon} \right) \right] \sigma_\mu^2.$$

As is evident, the variance of the optimal level of output depends positively on the variance of the real shock  $\sigma_\mu^2$ , and negatively on the variance of the monetary shock  $\sigma_\delta^2$ .

Since a rise in the variance of the real shock raises the value of  $\psi$ , its effect on the variance of optimal output is being magnified and the elasticity of  $\sigma_{y^*}^2$  with respect to  $\sigma_\mu^2$  exceeds unity.

### V. Constrained Optimization and Welfare

The analysis up to this point determined the optimal degrees of wage indexation and the optimal values of the response coefficients that govern monetary policy. The optimal values of the response coefficients were determined by minimizing the loss function. In this section we compute the values of the loss function that result from the adoption of various policy rules. This procedure enables us to compare the welfare loss that results from the imposition of alternative constraints on the degree of wage indexation, exchange rate intervention, and other policy instruments. The analysis also illustrates the more general proposition concerning the link between the information set and the number of independent response coefficients necessary for welfare maximization.

Using equations (32)–(33), the loss function can be written as

$$(44) \quad H = \phi^2 \sigma_\theta^2 - 2\phi(\sigma/(\sigma + \epsilon)) \times \psi(1 + \beta\sigma\psi) \sigma_{\mu-\delta}^2 + (\sigma^2 \psi^2 / (\sigma + \epsilon)^2) \sigma_{\mu-\delta}^2$$

$$\text{where } \sigma_\theta^2 = (\alpha - \tau)^2 \sigma_\rho^2 + (\alpha + \gamma - \xi)^2 \sigma_\chi^2 + (1 + \beta\sigma\psi)^2 \sigma_{\mu-\delta}^2.$$

We first consider the situation in which the only instrument of policy that can be set at its optimal level is the coefficient of wage indexation. In order to find the optimal value of the indexation coefficient, we note that in the loss function (44),  $b$  appears only in  $\phi$ ; therefore, minimization of  $H$  with respect to  $b$  is equivalent to minimization with respect to  $\phi$  (holding  $\gamma$  constant). This procedure yields the optimal value of  $\phi$ :

$$(45) \quad \phi^* = (\sigma/(\sigma + \epsilon))(1 + \beta\sigma\psi)(\sigma_\mu^2/\sigma_\theta^2),$$



By equating  $\phi^*$  with the definition of  $\phi$  in (33) we can obtain the optimal value of the indexation coefficient.

Substituting  $\phi^*$  for  $\phi$  in equation (44) and assuming that  $\xi = \tau = 0$ , the loss function becomes

$$(46) \quad H(b^*; \gamma) = \left( \sigma^2 \psi \sigma_\mu^2 / (\sigma + \varepsilon)^2 \right) \times \left[ \frac{\alpha^2 \sigma_\rho^2 + (\alpha + \gamma)^2 \sigma_\chi^2}{\alpha^2 \sigma_\rho^2 + (\alpha + \gamma)^2 \sigma_\chi^2 + (1 + \beta \sigma \psi)^2 \sigma_{\mu-\delta}^2} \right],$$

where  $H(b^*; \gamma)$  indicates that the loss function is evaluated under the condition that only the coefficient of wage indexation is set optimally, while the value of  $\gamma$  is set at an arbitrary level. When the exchange rate is fixed ( $\gamma = \infty$ ) the value of the loss function is

$$(46') \quad H(b^*, \gamma)|_{\gamma=\infty} = \left( \sigma^2 \psi \sigma_\mu^2 / (\sigma + \varepsilon)^2 \right),$$

and, when the exchange rate is flexible ( $\gamma = 0$ ) the value of the loss function is

$$(46'') \quad H(b^*, \gamma)|_{\gamma=0} = \left( \sigma^2 \psi \sigma_\mu^2 / (\sigma + \varepsilon)^2 \right) \times \left[ \frac{\alpha^2 (\sigma_\rho^2 + \sigma_\chi^2)}{\alpha^2 (\sigma_\rho^2 + \sigma_\chi^2) + (1 + \beta \sigma \psi)^2 \sigma_{\mu-\delta}^2} \right].$$

As is evident from comparison of equations (46') and (46''),

$$(47) \quad H(b^*, \gamma)|_{\gamma=\infty} \geq H(b^*, \gamma)|_{\gamma=0}.$$

Thus, except for extreme cases (like, for example, when there are no real shocks), the welfare loss for an economy for which only the wage indexation coefficient is set optimally is higher under fixed exchange rates than under flexible exchange rates. This result confirms the proposition established by Flood and Marion.

The forgoing analysis presumed that the value of  $\gamma$  is set at an arbitrary level that may not correspond to its optimal value. In order to obtain the optimal value of the coefficient of intervention in the foreign exchange market, we differentiate the loss func-

tion (44) with respect to  $\gamma$  and equate the derivative to zero:

$$(48) \quad \frac{\partial H}{\partial \phi} \frac{\partial \phi}{\partial \gamma} + 2(\alpha + \gamma - \xi) \sigma_\chi^2 \phi^2 = 0.$$

The assumption that the coefficient of wage indexation is set at its optimal value  $b^*$ , implies that at this point  $\partial H / \partial \phi = 0$  and, therefore, equation (48) implies that the optimal foreign exchange intervention coefficient is

$$(49) \quad \gamma^* = \xi - \alpha.$$

Substituting (49) into the loss function (44) and recalling that in the present stage of the analysis we have assumed that policy is constrained to set  $\xi = 0$ , yields

$$(50) \quad H(b^*, \gamma^*) = \left( \sigma^2 \psi \sigma_\mu^2 / (\sigma + \varepsilon)^2 \right) \times \left[ \frac{\alpha^2 \sigma_\rho^2}{\alpha^2 \sigma_\rho^2 + (1 + \beta \sigma \psi)^2 \sigma_{\mu-\delta}^2} \right],$$

where  $H(b^*, \gamma^*)$  indicates that the loss function is evaluated under the conditions that both wage indexation and exchange rate intervention are optimal. By subtracting equation (50) from (46) we obtain the marginal benefit from allowing optimal response to exchange rate changes. It can be shown that this marginal benefit is proportional to  $(\alpha + \gamma)^2 \sigma_\chi^2$ .

Equation (50) suggests that when  $\sigma_\rho^2 = 0$ ,  $H(b^*, \gamma^*) = 0$ . Thus, in this case, the optimal use of wage indexation and exchange rate intervention *eliminate* completely the welfare loss even though the value of  $\xi$  was constrained to equal zero. Likewise, inspection of equation (46) suggests that if both  $\sigma_\rho^2$  and  $\sigma_\chi^2$  are zero, as would be the case in a closed economy, then  $H(b^*) = 0$ . In this closed-economy case,  $b^* = b_c^*$  as in equation (38'), and the optimal use of the *single* instrument of wage indexation is capable of eliminating the welfare cost of labor market distortion.<sup>15</sup>

<sup>15</sup>A comparison between this result and that of Gray (1976) illustrates the role of the number of sources of

The economic intuition underlying these results can be stated in terms of the relation between the number of independent sources of information and the number of independent indicators governing policy rules. Two of the key assumptions underlying the model are that the level of employment is determined by the demand for labor and that real wages are adjusted according to an indexation formula that links the real wage *only* to the observed price level. The use of the price level in the adjustment of real wages as the only indicator to which the indexation rule applies may not permit an efficient use of the information that is available to economic agents. For example, in our model it is assumed that at each point in time individuals observe (or are able to infer without error) the shocks to prices  $\chi$ , the shocks to the rate of interest  $\rho$ , and the difference between the real and the monetary shocks,  $\mu - \delta$ . Adopting a single policy rule that links the real wage to the price level through the indexation coefficient may not use efficiently the more detailed information that is available in the open economy and that could be exploited in the adjustment of real wages.

---

information. In Gray's model there are two independent sources of information that are used in determining the equilibrium real wage. Therefore, the use of simple wage indexation *alone* does not eliminate the welfare loss. In contrast, when the present model is reduced to its closed-economy counterpart, there is only one independent source of relevant information (information about  $\mu - \delta$ ) and, therefore, the optimal use of the simple wage indexation rule eliminates the welfare loss. This discussion implies that if the magnitude of  $\mu$  were also known along with the knowledge of prices, interest rates, and the exchange rate, then the specification of the optimal money supply, that aims at eliminating the welfare loss, would include  $\mu$  as an additional indicator. In principle we could also introduce asymmetric information by which each individual firm observes their firm-specific shock but the aggregate economywide shock is not observable; for related specifications, see Flood-Hodrick and Marston-Turnovsky (1983b). It is important to emphasize that in our framework the welfare cost does not reflect assumptions concerning incomplete or asymmetric information. Rather, the information set is assumed to be equally shared among market participants and policymakers but, due to contracts, this shared information may not be immediately acted upon by market participants.

This is the reason for the proposition that, except for special cases,  $H(b^*, \gamma) > 0$ .

In equation (46), the value of the term in the squared brackets characterizes the quality of the use of information in the adjustment of real wages. When this term is zero, as would for example be the case when  $\sigma_p^2 = \sigma_\chi^2 = 0$ , then the information set is used most efficiently in the sense that the observed price provides all of the relevant information for determining the optimal adjustment of real wages. Under such circumstances, indeed, the optimal indexation coefficient  $b^*$  eliminates the welfare loss, as would be the case in the closed economy. In general, however, if  $\sigma_p^2$  or  $\sigma_\chi^2$  are positive, then the squared bracket term in equation (46) is positive, indicating that the adoption of a single policy indicator when there are more independent sources of information does not result in a market-clearing real wage and, therefore, does not eliminate the welfare loss. Another illustration of this argument is provided by equation (50) where it is assumed that the policy is governed by two independent indicators. Under such circumstances, if there are three independent sources of information (the observed values of  $s$ ,  $p$ , and  $\rho$ ), the adoption of the optimal values  $b^*$  and  $\gamma^*$  does not eliminate the welfare loss and  $H(b^*, \gamma^*) > 0$ . In contrast, if there were no shocks to the rate of interest, there would only be two independent sources of information ( $s$  and  $p$ ); in such a case the term in the squared brackets in equation (50) would be zero, indicating that the optimal response to the two indicators is capable of eliminating the welfare loss since it generates the market-clearing real wage.

The forgoing discussion dealt with the policies necessary for the elimination of the welfare loss arising from suboptimal real wages. The fundamental proposition, however, is more general. Policies can be designed to eliminate the welfare cost of distortions. The general principle developed by Tinbergen states that in order to attain  $n$  targets, economic policy must possess at least  $n$  independent instruments. Our application demonstrated that with the necessary number of instruments, the optimal policy will

succeed in attaining the targets only if the instruments are influenced by a sufficient number of independent indicators.<sup>16</sup> This sufficient number must equal the number of independent sources of information that influence the determination of the undistorted level of the targets.

## VI. Concluding Remarks

In this paper we have analyzed the relation between the optimal degrees of wage indexation and foreign exchange intervention. The optimal values of these policy instruments were obtained as components of the solution to the broader problem of the design of optimal monetary policy. The model used for the analysis was governed by the characteristics of the stochastic shocks that affect the economy and by the information set that individuals were assumed to possess.

Throughout the analysis the optimal policies are obtained with reference to an objective function that has the desirable property of possessing explicit welfare justification. It represents the welfare loss arising from the assumptions that employment is governed by the demand for labor, nominal wages are precontracted, and real wages are adjusted according to an indexation formula that links the real wage to the observed price. The use of the price level in the adjustment of real wages as the only indicator to which the indexation rule applies may not permit an efficient use of the information that is available to economic agents and that could be exploited in the adjustment of real wages. The loss function reflects the welfare cost associated with a discrepancy between the equilibrium change in real wages that would occur under an optimal use of information and the actual change in real wages that

results from labor market conventions and from the adoption of specific policy rules.

One of the key findings of the paper concerns the conditions under which the optimal policy, by minimizing the loss function, also *eliminates* the welfare cost. It was shown that if the number of independent indicators that govern policy is equal to the number of independent sources of information that are relevant for the determination of the market-clearing real wage, then the adoption of the optimal policy rules eliminates completely the welfare cost of labor market distortions. This proposition is important, since the elimination of the welfare cost implies that the optimal policies are capable of reproducing the equilibrium that would be obtained under the assumption that labor markets were cleared after the realization of the stochastic shocks. By reproducing that equilibrium, the optimal policies nullify the distortions that result from the assumption that, because of contracts, nominal wages are predetermined. When such an optimum obtains, the important issues concerning the implications of the assumption that employment is determined by the demand for labor, as raised by Cukierman, are inconsequential since, at the optimum, there is an equality between the demand and the supply of labor. Similarly, when the optimum obtains many of the critical issues concerning the conceptual difficulties associated with the existence of suboptimal contracts, as raised by Barro, are also inconsequential since, at the optimum, the contracts (along with the optimal policy) are optimal. In that sense the equilibrium which eliminates the welfare loss is analogous to the closed-economy equilibria that were analyzed by Karni and Aizenman (1983).

The principle underlying the determination of the optimal set of policy rules was illustrated in terms of the design of a sophisticated monetary policy. Alternatively, in situations where a sophisticated monetary policy is not feasible, analogous considerations could be incorporated into the design of a sophisticated wage indexation formula. As long as each independent source of information that is relevant for the determination

<sup>16</sup> The requirement that the indicators must be independent is reflected in our case by the exclusion of the price  $p$  from the set of indicators governing the supply of money. Clearly, of the triplet  $p, s, \chi$ , only two contain independent information that can be usefully exploited. Thus, of the three, we chose to include  $s$  and  $\chi$  in the set of indicators.

of the market-clearing real wage serves as an independent indicator for policy that is used optimally, the resulting equilibrium replicates the distortion-free equilibrium.

Our analysis showed that when wage indexation serves as one of the independent policy rules, then a rise in the variance of the real productivity shock and a rise in the elasticity of output with respect to labor input lower the optimal degree of wage indexation. On the other hand, a rise in the variance of the monetary shock and a rise in the elasticity of labor supply raise the optimal degree of wage indexation. It was also shown that when the degree of foreign exchange intervention is exogenously given, then, around the optimum, a rise in the degree of exchange rate flexibility raises the optimal degree of wage indexation. Likewise, when the degree of wage indexation is exogenously given, then, around the optimum, a higher degree of wage indexation raises the optimal degree of exchange rate flexibility.

We concluded our discussion with an examination of the consequences of departures from optimal policies. In this context we compared the welfare loss that results from the imposition of alternative constraints on the degree of wage indexation, on foreign exchange intervention and on the magnitudes of other policy coefficients.

One of the limitations of the analysis in this paper relates to the level of aggregation. We have assumed that there is one composite good which is internationally traded at a (stochastically) given world price. A useful extension would allow for a richer menu of commodities including those that are internationally tradable and those which are nontradable. The presence of nontradable goods would then relax some of the constraints that were imposed by the small country assumption. Owing to its relative size, the economy would still be a price taker in the world traded-goods market, but the relative price of its nontraded goods would be endogenously determined by market-clearing conditions. Such an extension should facilitate the distinction between mechanisms and policies that operate on the price level and those that operate on relative prices. It would allow for a distinction between the relative price of

traded goods—the external terms of trade—and the relative price of nontraded goods—the internal terms of trade (the real exchange rate)—and would facilitate an analysis of deviations from purchasing power parity. The introduction of nontraded goods should also permit an analysis of the influence of the degree of openness (as measured by the relative size of the traded-goods sector) on the optimal values of the coefficients that govern policy. Previous studies suggest that the degree of openness may play a significant role in influencing optimal policies (our 1982 paper and Aizenman, 1985). The broader menu of goods should also facilitate an analysis of the optimal indexation rules in the face of supply shocks (as in Marston and Turnovsky, 1983a), as well as an analysis of the proper price index that should be used in the indexation formula (as in Marston, 1982b, and in our 1985 article).

Another extension of the analysis would draw an explicit distinction between permanent and transitory shocks. In our specifications the stochastic disturbances were assumed to be independent of each other and to be drawn from a distribution with a constant variance and a zero mean. A more complete analysis would distinguish between permanent and transitory shocks and would incorporate the role of time preference; thereby, it would introduce dynamic considerations into the analysis of the optimal choice of policies.

It is relevant to note that the nature of labor contracts assumed in this paper was motivated by realism. Accordingly, we assumed that contracts specify the nominal wage whereas the level of output is determined by firms, and that the indexation formula is simple in that it adjusts wages to changes in the price level rather than to a complex set of variables. Our analysis does not attempt to contribute to the theory that explains this conventional form of labor contract (on this see Barro and Fischer, 1977b).

Finally, we define the equilibrium that replicates the performance of an economy in which labor markets clear without friction, as the social optimum. Implicit in this definition of the social optimum is the assumption

that individuals and firms are risk neutral since, in general (as shown by Costas Azariadis, 1978), when attitudes towards risk differ across economic agents, auction markets do not allocate risk efficiently and individuals find it advantageous to enter into long-term risk-sharing contracts. Our assumption, therefore, precludes rationalizing the existence of labor contracts in terms of the insurance function. Therefore, in this framework (as in Gray, 1978), the existence of contracts reflects the cost of negotiations.

## APPENDIX

### *The Derivation of the Objective Function*

In the two parts of this Appendix we provide a formal justification for the specification of the loss function.

*Part I:* Define by  $(\bar{W}/P)$  the equilibrium real wage that clears the labor market. This equilibrium value of the real wage clears the market for any given expected value of the real shock conditional on the available information. Since in equilibrium the real wage equals the expected marginal product of labor, the amount of labor,  $\bar{L}$ , that clears the labor market when the real wage is  $(\bar{W}/P)$  is defined by

$$(A1) \quad E[Y_L(\bar{L})|I_t] = (\bar{W}/P),$$

where  $I_t$  denotes the information set available at time  $t$ , and where  $Y_L(\bar{L})$  denotes the marginal product of labor evaluated at  $L = \bar{L}$ . For subsequent use it is convenient to define the function  $\bar{X}$  as the expected value of  $X$  conditional on the available information  $I_t$ . Thus, applying this notation to equation (A1) yields:

$$(A1') \quad \bar{Y}_L(\bar{L}) = (\bar{W}/P).$$

General equilibrium requires that the level of employment  $L$  be also consistent with the supply of labor that is supplied by utility-maximizing workers at the given real wage  $(\bar{W}/P)$ . To illustrate, let the utility func-

tion be

$$(A2) \quad u(C, L); \quad \partial u / \partial C > 0, \quad \partial u / \partial L < 0,$$

where  $C$  denotes the level of consumption and  $L$  denotes labor, that is, negative leisure. Maximization of the utility function subject to the technological constraint that production  $Y$  is governed by the production function  $Y = F(L)$ , and that, from the budget constraint in the absence of asset accumulation, the values of production and consumption must coincide, yields the desired supply of labor. In general equilibrium, with  $L = \bar{L}$ , the equilibrium level of utility is denoted by  $U(\bar{L})$ .

In practice, due to a precontracted nominal wage, the realized real wage may differ from its full equilibrium level. Since by assumption employment is demand determined, it follows that the actual level of employment,  $L$ , when the real wage differs from  $(\bar{W}/P)$ , differs from  $\bar{L}$  and, associated with this level of employment and production, the level of utility is  $U(L)$ .

The welfare cost of suboptimal employment is  $[U(\bar{L}) - U(L)]/\lambda$  where  $\lambda$  measures the marginal utility of income. Using Arnold Harberger's formulation (1971, equation 5') for the analysis of consumer's surplus, we expand the utility function in a Taylor series around the general equilibrium and omit third-order terms to obtain Harberger's expression for the approximation of the welfare loss:

$$(A3) \quad \Delta U / \lambda \approx \sum [P_i^0 \Delta C_i - \frac{1}{2} \Delta P_i \Delta C_i],$$

where  $\Delta C_i$  denotes the change in the rate of consumption of good  $i$ ,  $P_i^0$  denotes the equilibrium price of good  $i$ , and where  $\Delta P_i$  measures the discrepancy of the full equilibrium price from the actual price. In applying (A3) to the utility function assumed here, it is useful to decompose the expression into terms involving goods and those involving labor (or leisure). In our case, with a single (composite) commodity which is used as the numeraire, an application of Harberger's formula yields  $\Delta C$  as the welfare change associated with the change in the consumption of

that good. The same procedure is also applied to labor, which is the second argument in the utility function  $u(C, L)$ , and whose equilibrium price is  $-(\bar{W}/P)$ . For that component we obtain  $-(\bar{W}/P)\Delta L + \frac{1}{2}\Delta(\bar{W}/P)^s\Delta L$ , where the change in the real wage is measured along the supply of labor that reflects the utility function. Combining the expressions measuring the welfare cost of changes in consumption and labor yields (A4) as the welfare loss:

$$(A4) \quad \Delta U/\lambda \approx \Delta C - (\bar{W}/P)\Delta L + \frac{1}{2}\Delta(\bar{W}/P)^s\Delta L,$$

where  $\Delta L = \tilde{L} - L$ . In computing the value of the expression in (A4) we simplify the specification of the intertemporal budget constraint by specifying a temporal budget constraint according to which  $\Delta C = \Delta Y$  and, therefore, the value of  $\Delta C$  can be obtained by calculating  $\Delta Y$ . It is noteworthy, however, that in general, the values of expenditure and income in each period need not be equal to each other as long as there is equality between their discounted present values. With undistorted capital markets, however, the two formulations are equivalent as is known from the literature on the Ricardian Equivalence. In Part II of this Appendix we discuss the intertemporal model in greater detail. Expanding the production function in a Taylor series around the general equilibrium up to the second-order terms yield

$$(A5) \quad \Delta Y = \bar{Y}_L\Delta L - \frac{1}{2}\bar{Y}_{LL}(\Delta L)^2.$$

where  $\Delta Y = \bar{Y}(\tilde{L}) - Y(L)$ . Since the expansion is around the equilibrium, we substitute in (A5) the equilibrium real wage for  $\bar{Y}_L$ , and then substitute the resulting expression for  $\Delta C (= \Delta Y)$  into (A4):

$$(A6) \quad \Delta U/\lambda = -\frac{1}{2}\bar{Y}_{LL}(\Delta L)^2 + \frac{1}{2}\Delta(\bar{W}/P)^s\Delta L,$$

or, equivalently,

$$(A6') \quad \Delta U/\lambda = \frac{1}{2}(-\bar{Y}_{LL} + \epsilon')(\Delta L)^2,$$

where  $\epsilon'$  denotes the slope of the supply of labor, that is,  $\epsilon' \equiv \Delta(\bar{W}/P)^s/\Delta L$ . Multiplying and dividing the right hand side of (A6') by  $\tilde{L}^2$  yields

$$(A6'') \quad \Delta U/\lambda = (\tilde{L}^2/2)[- \bar{Y}_{LL} + \epsilon'](\tilde{l} - l)^2,$$

where  $\tilde{l} - l$ , which is defined as  $\log(\tilde{L}/L_0) - \log(L/L_0)$ , approximates the percentage discrepancy between equilibrium and actual employment levels (where we employ the approximation that  $\log(1+x) \approx x$ ). Using the production function to compute  $\bar{Y}_{LL}$ , we obtain

$$(A7) \quad \Delta U/\lambda = \left(\frac{1}{\epsilon} + \frac{1}{\sigma}\right)\frac{\tilde{L}}{2}\left(\frac{\tilde{W}}{P}\right)(l - \tilde{l})^2,$$

where  $\epsilon$  denotes the elasticity of labor supply and, as defined in the text,  $\sigma = 1/(1-\beta)$  where  $\beta$  denotes the elasticity of output with respect to labor input. Substituting for  $(l - \tilde{l})^2$  from equation (27) in the text yields

$$(A8) \quad \frac{\Delta U}{\lambda} = \frac{\sigma^2}{2}\left(\frac{1}{\epsilon} + \frac{1}{\sigma}\right) \times \left[-(w - p) + \frac{\sigma E(\mu)}{\sigma + \epsilon}\right]^2 \left(\frac{\tilde{W}}{P}\right)\tilde{L},$$

where (A8) corresponds to equation (30) in the text. Our loss function is defined as the expected value of the welfare loss from sub-optimal employment during period  $t$  resulting from the existence of contracts that were agreed upon on the basis of information available at period  $t-1$ . Thus, ignoring the constant  $(\sigma^2/2)(1/\epsilon + 1/\sigma)$  and treating the equilibrium wage bill,  $(\bar{W}/P)\tilde{L}$ , as constant (i.e., ignoring third-order terms of Taylor expansion) we obtain the loss function  $H$ :

$$(A9) \quad H = E\left[\left\{-(w - p) + \frac{\sigma}{\sigma + \epsilon}E(\mu)\right\}^2 | I_{t-1}\right],$$

which corresponds to equation (31) in the text.

*Part II:* Here we demonstrate that the same objective function also applies to a dynamic

intertemporal model. Consider a two-period model and let the present value of utility  $U$  be

$$(A10) \quad U = u(C_1, L_1) + \bar{p}u(C_2, L_2),$$

where  $\bar{p}$  designates the subjective discount factor and where the subscripts 1 and 2 designate variables pertaining to periods 1 and 2, respectively. The value of assets which are not consumed in period 1 is  $A_1$  and their value in period 2 is  $(1+r)A_1$  where  $r$  designates the exogenously given (stochastic) world rate of interest on internationally traded bonds. Denoting rents by  $R$  and assuming that these are redistributed as lump sum transfers, the periodic transfers are the value of output minus the wage bill:

$$(A11) \quad R_t = Y_t(L_t) - \left(\frac{W}{P}\right)_t L_t, \quad (t=1,2).$$

The formal maximization problem in period 2 can be written as

$$(A12) \quad \max E_2[u(C_2, L_2)]$$

subject to

$$C_2 = (1+r)A_1 + \left(\frac{W}{P}\right)_2 L_2 + R_2.$$

The solution to the maximization problem in (A12) yields the optimal values of consumption and labor supply in that period,  $C_2^*$  and  $L_2^*$ . These optimal values are conditional, of course, on the historically given value of  $A_1$ . Thus, we can define a function  $u^*(A_1)$  where  $u^*(A_1) \equiv u(C_2^*, L_2^*)$ . We then can present the maximization problem for period 1 as

$$(A13) \quad \max E_1[u(C_1, L_1) + \bar{p}u^*(A_1)],$$

$$\text{subject to } C_1 = Q + \left(\frac{W}{P}\right)_1 L_1 + R_1 - A_1,$$

where  $Q$  denotes the given initial endowment. The solution to (A13) yields the optimal values  $\bar{C}_1$ ,  $\bar{L}_1$ , and  $\bar{A}_1$ . As before we denote the value of utility in the general equilibrium by  $U(\bar{L}_1)$  where it is understood

that this level of utility is obtained when  $C_1$ ,  $L_1$ , and  $A_1$  are set at their unconstrained optimal values  $\bar{C}_1$ ,  $\bar{L}_1$ , and  $\bar{A}_1$ . In practice, due to contracts, the level of employment may be constrained to  $L_1$ . The resulting level of utility would be  $U(L_1)$ , where it is understood that  $C_1$  and  $A_1$  are still chosen optimally subject to the constraint that the labor supply is  $L_1$ . The welfare cost would be  $[U(\bar{L}_1) - U(L_1)]/\lambda$ .

In computing the welfare cost we apply Harberger's formula (A3) where we note that in the present case one of the arguments in the underlying utility function is the asset position  $A_1$ . Thus the expression corresponding to (A4) becomes

$$(A14) \quad \Delta U/\lambda \approx \Delta C_1 - (\bar{W}/\bar{P})_1 \Delta L_1 + \Delta A_1 + \frac{1}{2} \Delta (W/P)_1^s \Delta L_1.$$

Using the constraint in (A13) we obtain

$$(A15) \quad \Delta C_1 + \Delta A_1 = \Delta \left[ \left(\frac{W}{P}\right)_1 L_1 + R_1 \right],$$

and using the definition of  $R_1$  from (A11) the expression in (A15) equals the change in output  $\Delta Y_1$ , which is substituted in (A14) to yield

$$(A16) \quad \Delta U/\lambda \approx \Delta Y_1 - (\bar{W}/\bar{P})_1 \Delta L_1 + \frac{1}{2} \Delta (W/P)_1^s \Delta L_1.$$

Finally, by using (A5) for  $\Delta Y_1$  we can proceed along identical lines to those in Part I of this Appendix so as to yield the objective function used in the text.

We have thus demonstrated that the dynamic problem could be transformed into the simpler static counterpart used in Part I of the Appendix. The assumption necessary for this transformation is that the utility function is separable. To ensure that the resultant objective function remains invariant with respect to monetary policy, it is also assumed that the private sector fully internalizes the central bank's asset position.

Prior to concluding this Appendix we can now use the loss function in order to justify

our earlier assumption concerning the initial nominal wage,  $W_0$ . Using the definition of the variance, the loss function (A9) can be written as

$$(A9') \quad H = \text{Var} \left[ -(w-p) + \frac{\sigma}{\sigma + \varepsilon} E(\mu) | I_{t-1} \right] \\ + E \left[ \left\{ -(w-p) + \frac{\sigma}{\sigma + \varepsilon} E(\mu) \right\} | I_{t-1} \right]^2.$$

Our assumption that the initial contractual nominal wage is set at the level that would have prevailed in equilibrium in the absence of shocks, is necessary in order to ensure that the second term on the right-hand side of (A9') vanishes; for any other choice of  $W_0$  this second term would be positive (since  $E[-(w-p) | I_{t-1}]$  would not be zero), and the welfare loss would not be minimized.

#### REFERENCES

- Aizenman, Joshua, "Wage Contracts with Incomplete and Costly Information," NBER Working Paper Series, No. 1150, June 1983.
- , "Wage Flexibility and Openness," *Quarterly Journal of Economics*, 1985 forthcoming.
- and Frenkel, Jacob A., "Supply Shocks and Optimal Wage Indexation in the Open Economy," in J. Edwards and L. Ahamed, eds., *Structural Adjustment and the Real Exchange Rate in Developing Countries*, Chicago: University of Chicago Press, 1985 forthcoming.
- Azariadis, Costas, "Escalator Clauses and the Allocation of Cyclical Risks," *Journal of Economic Theory*, June 1978, 18, 119–55.
- Barro, Robert J., "Long-Term Contracting, Sticky Prices, and Monetary Policy," *Journal of Monetary Economics*, July 1977, 3, 305–16.
- Bhandari, Jagdeep S., "Staggered Wage Setting and Exchange Rate Policy in an Economy with Capital Assets," *Journal of International Money and Finance*, December 1982, 1, 275–92.
- Calvo, Guillermo A., "On Models of Money and Perfect Foresight," *International Economic Review*, February 1979, 20, 83–102.
- Canzoneri, Matthew B., Henderson, Dale W. and Rogoff, Kenneth S., "The Information Content of the Interest Rate and Optimal Monetary Policy," *Quarterly Journal of Economics*, November 1983, 98, 545–66.
- Cukierman, Alex, "The Effects of Wage Indexation on Macroeconomic Fluctuations: A Generalization," *Journal of Monetary Economics*, April 1980, 6, 147–70.
- Fischer, Stanley, (1977a) "Wage Indexation and Macroeconomic Stability," in Karl Brunner and Allan H. Meltzer, eds., *Stabilization of Domestic and International Economy*, Vol. 5, Carnegie-Rochester Conferences on Public Policy, *Journal of Monetary Economics*, Suppl. 1977, 107–47.
- , (1977b) "Long-Term Contracting, Sticky Prices, and Monetary Policy: A Comment," *Journal of Monetary Economics*, July 1977, 3, 317–23.
- Flood, Robert P. and Marion, Nancy P., "The Transmission of Disturbance under Alternative Exchange-Rate Regimes with Optimal Indexation," *Quarterly Journal of Economics*, February 1982, 97, 43–66.
- and Hodrick, Robert J., "Optimal Price and Inventory Adjustment in an Open Economy Model of the Business Cycle," NBER Working Paper Series, No. 1089, March 1983.
- Frenkel, Jacob A. and Aizenman, Joshua, "Aspects of the Optimal Management of Exchange Rates," *Journal of International Economics*, November 1982, 13, 231–56.
- and Razin, Assaf, "Stochastic Prices and Tests of Efficiency of Foreign Exchange Markets," *Economics Letters*, No. 2, 1980, 6, 165–70.
- Gray, Jo Anna, "Wage Indexation: A Macroeconomic Approach," *Journal of Monetary Economics*, April 1976, 2, 221–35.
- , "On Indexation and Contract Length," *Journal of Political Economy*, February 1978, 86, 1–18.
- Harberger, Arnold C., "Three Basic Postulates for Applied Welfare Economics: An Interpretive Essay," *Journal of Economic Literature*, September 1971, 9, 785–97.
- Karni, Edi, "On Optimal Wage Indexation," *Journal of Political Economy*, April 1983, 91, 282–92.



- McCallum, Bennett T., "On Non-Uniqueness in Rational Expectations Models: An Attempt at Perspective," *Journal of Monetary Economics*, March 1983, 11, 139-68.
- Marston, Richard C., (1982a) "Wages, Relative Prices and the Choice Between Fixed and Flexible Exchange Rates," *Canadian Journal of Economics*, February 1982, 15, 87-103.
- \_\_\_\_\_, (1982b) "Real Wages and the Terms of Trade: Alternative Indexation Rules for an Open Economy," NBER Working Paper Series, No. 1046, December 1982.
- \_\_\_\_\_ and Turnovsky, Stephen J., (1983a) "Imported Material Prices, Wage Policy and Macroeconomic Stabilization," NBER Working Paper Series, No. 1254, December 1983.
- \_\_\_\_\_ and \_\_\_\_\_, (1983b) "Macroeconomic Stabilization Through Taxation and Indexation: The Use of Firm Specific Information," unpublished manuscript, December 1983.
- Sachs, Jeffrey, "Wages, Flexible Exchange Rates, and Macroeconomic Policy," *Quarterly Journal of Economics*, June 1980, 94, 731-47.
- Turnovsky, Stephen J., (1983a) "Wage Indexation and Exchange Market Intervention in a Small Open Economy," *Canadian Journal of Economics*, November 1983, 16, 574-92.
- \_\_\_\_\_, (1983b) "Exchange Market Intervention Policies in a Small Open Economy," in J. Bhandari and B. Putman, eds., *Economic Interdependence and Flexible Exchange Rates*, Cambridge: MIT Press, 1983, 286-311.

# Network Externalities, Competition, and Compatibility

By MICHAEL L. KATZ AND CARL SHAPIRO\*

There are many products for which the utility that a user derives from consumption of the good increases with the number of other agents consuming the good. There are several possible sources of these positive consumption externalities.<sup>1</sup>

1) The consumption externalities may be generated through a direct physical effect of the number of purchasers on the quality of the product. The utility that a consumer derives from purchasing a telephone, for example, clearly depends on the number of other households or businesses that have joined the telephone network. These network externalities are present for other communications technologies as well, including Telex, data networks, and over-the-phone facsimile equipment.

2) There may be indirect effects that give rise to consumption externalities. For example, an agent purchasing a personal computer will be concerned with the number of other agents purchasing similar hardware because the amount and variety of software that will be supplied for use with a given computer will be an increasing function of the number of hardware units that have been sold. This hardware-software paradigm also applies to video games, video players and recorders, and phonograph equipment.

3) Positive consumption externalities arise for a durable good when the quality and availability of postpurchase service for the good depend on the experience and size of the service network, which may in turn vary with the number of units of the good that have been sold. In the automobile market, for example, foreign manufacturers' sales initially were retarded by consumers' awareness of the less experienced and thinner service networks that existed for new or less popular brands.

In all of these cases, the utility that a given user derives from the good depends upon the number of other users who are in the same "network" as is he or she. The scope of the network that gives rise to the consumption externalities will vary across markets. In some cases, such as the automobile example, the sales of only one firm will constitute the relevant network. In other cases, the relevant network will comprise the outputs of all firms producing the good. For example, the number of stereo phonographs of any one brand is not a determinant of the supply of records that a consumer can play on his or her stereo. In still other markets, the network may comprise the products of a coalition of firms that is a subset of the entire market, as in the case of computers, where some groups of manufacturers adopt common operating systems.

The central feature of the market that determines the scope of the relevant network is whether the products of different firms may be used together. For communications networks, the question is one of whether consumers using one firm's facilities can contact consumers who subscribe to the services of other firms. If two firms' systems are interlinked, or compatible, then the aggregate number of subscribers to the two systems constitutes the appropriate network. If the systems are incompatible, such as Telex and cable, then the size of an individual

\*Department of Economics and Woodrow Wilson School, respectively, Princeton University, Princeton, NJ 08544. We thank the anonymous referees for helpful suggestions. We also thank seminar participants at Princeton, Columbia, Queens, the Federal Trade Commission, and MIT for comments. This material is based upon work supported by the National Science Foundation under grants nos. SES-820942 and SES-82-07377.

<sup>1</sup>In addition to the sources of consumption externalities mentioned in this paper, there are a number of more subtle ones. These include: (i) the fact that product information is more easily available for more popular brands; (ii) the role of market share as a signal of product quality; and (iii) purely psychological, bandwagon effects.

system is the proper network measure for users of that system.

Similarly, for hardware-software markets, the issue is whether software produced for use on one brand of hardware may be run on another brand of hardware. If two brands of hardware can use the same software, then the hardware brands are said to be compatible. The relevant network is the set of users who have compatible brands of hardware. In the personal computer market, the CPM operating system has been designed to allow several brands of computers to use common programs. In the case of quadraphonic audio discs, on the other hand, the records made for one type of player cannot be used on a player that uses a different quadraphonic technology. Here, unlike the case of stereos, the relevant network for a given brand of equipment comprises the set of brands that use the same technology, not the entire market.

For the durables example, the relevant network is the set of brands that require the same parts of servicing skills. If a particular model of automobile has customized parts or requires specialized repair skills, then an owner of the model will find a thinner, and probably more expensive, service system. This smaller network will reduce his or her initial willingness to pay for the model.

Despite the significance of markets in which network externalities are present, relatively little work has been done in this area. The analysis done so far has been set in a monopoly context and has focussed on communications networks. Shmuel Oren and Stephen Smith (1981) is a recent reference to this literature. As the examples above make clear, it is important to extend the study of network externalities to an oligopolistic setting.

In this paper, we develop a simple, static model of oligopoly to analyze markets in which consumption externalities are present. We examine two basic sets of issues. First, we study the effect of consumption externalities on competition and the form of the market equilibrium. When network externalities exist, consumers must form expectations regarding the size of competing networks.

We use a notion of rational, or fulfilled expectations, equilibrium. Our basic findings are that consumption externalities give rise to demand-side economies of scale, which will vary with consumer expectations. As a result, multiple fulfilled expectations equilibria may exist for a given set of cost and utility functions. For some sets of expectations only one firm will produce output, while for other sets of expectations there will be several firms in the market. These equilibria verify the following intuition: if consumers expect a seller to be dominant, then consumers will be willing to pay more for the firm's product, and it will, in fact, be dominant.

The second area that we explore is the compatibility decision. Typically, firms can choose whether to manufacture compatible products, and thus can determine whether individual firm or aggregate market sales are the relevant ones in the evaluation of the consumption externalities. An important question, therefore, is whether firms will have proper incentives to produce compatible goods or services.

Gerald Brock (1975) and Robert Kurdle (1975) have done interesting case studies of compatibility decisions in the U.S. mainframe computer and farm machinery industries, respectively. Neither, however, develops a model of equilibrium in which to analyze firms' compatibility incentives. Using our model, we compare the private and social incentives to produce compatible products. We find that firms with good reputations or large existing networks will tend to be against compatibility, even when welfare is increased by the move to compatibility. In contrast, firms with small networks or weak reputations will tend to favor product compatibility, even in some cases where the social costs of compatibility outweigh the benefits. Viewing firms as a collective decision maker, we find that in our model the firms' joint incentives for product compatibility are lower than the social incentives.

The paper is organized as follows. In Section I, we present the model and define the equilibrium concept. The set of market equilibria is characterized in Section II. We ex-

amine the private and social incentives for compatibility in Section III, primarily under the assumptions that all of the costs of compatibility are fixed. There is a brief summary of our results and a discussion of the relevance of these results for public policy in Section IV. We outline an alternative approach to the formation of consumer expectations in the Appendix. It is shown that the equilibria are qualitatively similar if the firms are able to commit to given network sizes before consumers make their purchase decisions.

## I. A Formal Model of Network Competition

### A. Consumers

We look at a partial equilibrium oligopoly model in which there are no income effects and consumers act to maximize their surplus. A consumer buys at most one brand and purchases either one or no unit of any given brand.

The surplus that a consumer derives from buying a unit of the good depends on the number of other agents who join the network associated with that product. When the good is durable, an individual's consumption benefits will depend on the future size of the relevant network. Consumers will base their purchase decisions on *expected* network sizes. To capture this important feature of many markets with network externalities in our static, one-period model, we assume that consumers must make their purchase decisions before the actual network sizes are known. The timing is as follows. First, consumers form expectations about the size of the network with which each firm is associated. Second, the firms play an output game, taking consumers expectations as given. This game generates a set of prices. Consumers then make their purchase decisions by comparing their reservation prices (based on expected network sizes) with the prices set by the  $n$  firms,  $i = 1, \dots, n$ .

We do not explicitly model the process through which consumers' expectations are formed. We will, however, impose the requirement that in equilibrium consumers' expectations are fulfilled. Let  $x_i^e$  denote the

number of customers that a consumer expects firm  $i$  to have, and let  $y_i^e$  be the consumer's prediction of the size of the network with which firm  $i$  is associated. All consumers are assumed to have identical expectations of network sizes. When the brands are incompatible, each makes up its own network so  $y_i^e = x_i^e$ . When  $m$  firms' products are compatible, say brands 1 through  $m$ , then there is a single network for these brands and

$$y_i^e = \sum_{j=1}^m x_j^e \quad \text{for } i = 1, 2, \dots, m.$$

Networks are assumed to be homogeneous in the sense that if two networks are of equal size, then all consumers view the two networks as perfect substitutes.

Consumers are assumed to be heterogeneous in their basic willingness to pay for the product, but homogeneous in their valuations of the network externality. Specifically, a consumer of type  $r$  has a willingness to pay  $r + v(y^e)$  for a product with expected network size  $y^e$ . Without further loss of generality we can normalize  $r$  and  $v(0)$  so that  $v(0) = 0$ . We can interpret  $r$  as the consumer's basic willingness to pay for the good and  $v(y)$  as the value he or she attaches to the consumption externality when the number of subscribers is  $y$ . The externality function is taken to be twice continuously differentiable, with  $v' > 0$ ,  $v'' < 0$ , and  $\lim_{y \rightarrow \infty} v'(y) = 0$ . The basic willingness to pay for the good,  $r$ , varies across consumers and is assumed to be uniformly distributed between minus infinity and  $A$  with density one.<sup>2</sup> We assume that  $A$  is positive.

Each agent purchases the brand that maximizes his or her surplus. Letting  $p_i$  denote the price charged for brand  $i$ , a consumer of type  $r$  chooses the brand for which

$$(1) \quad r + v(y_i^e) - p_i$$

<sup>2</sup> The uniform density assumption amounts to assuming a linear demand curve for the product. We assume that the support of  $r$  has no finite lower limit in order to avoid having to consider corner solutions, where all consumers enter the market.

is largest. If (1) is negative for all  $i$ , then a type  $r$  consumer stays out of the market and purchases none of the brands.

### B. Firms

Given the homogeneity of the products, two firms  $i$  and  $j$  will both have positive sales only if

$$(2) \quad p_i - v(y_i^e) = p_j - v(y_j^e),$$

where  $p_i - v(y_i^e)$  is the expected hedonic price of brand  $i$ , that is, the price adjusted for the network size. Equation (2) says that the hedonic prices must be equal when multiple firms have positive sales. Let  $\phi$  denote the common value of the hedonic prices given in equation (2).

For a given value of  $\phi$ , only those consumers for whom  $r \geq \phi$  enter the market. Given the uniform distribution of  $r$ , there are  $A - \phi$  such consumers. Thus, if the firms sell a total of  $z \equiv \sum_{i=1}^n x_i$  units, then prices must be set such that  $A - \phi = z$ , or

$$(3) \quad A + v(y_i^e) - p_i = z$$

for all  $i$  such that  $x_i > 0$ .

From equation (3) we see that firm  $i$  receives a price of

$$(4) \quad p_i = A + v(y_i^e) - z.$$

The price that firm  $i$  receives depends on the expected size of its network,  $y_i^e$ , and on the total unit sales of the  $n$  firms,  $z$ .

There are two types of costs that must be modeled. First, there are costs of production. We assume that production costs are the same for all firms and that these costs take the form of a fixed cost,  $G$ , plus a constant per unit variable cost,  $g$ . That is, the cost to firm  $i$  of producing  $x$  units of output is  $G + gx$ . As long as the fixed costs are smaller than the firm's equilibrium revenues minus variable costs, the fixed costs have no effect on the equilibrium. To simplify the exposition, we assume that the fixed costs of production are equal to zero. Without loss of generality, we also take the variable costs of

production to be equal to zero. Assuming that  $g$  is equal to zero is equivalent to redefining  $r$  to be the excess of the consumer's basic willingness to pay for the good over the constant per unit cost.<sup>3</sup>

There is a second type of cost that we must consider, the cost of the achieving compatibility. For most of the analysis we will assume that the costs of compatibility are fixed costs, that is, are independent of scale. This amounts to assuming that compatible products have the same marginal production costs as incompatible products. (We discuss the consequences of relaxing this assumption of Section III, Part D). The fixed cost of compatibility that we analyze could include costs of developing and designing a compatible product, the costs of negotiating to select a standard, and the costs of introducing a new, compatible product. Let  $F_i$  denote the fixed costs of compatibility incurred by firm  $i$ . Note that  $F_i$  need not be the same for all firms.

If all of the networks are incompatible, then  $y_i^e = x_i^e$ , and firm  $i$  earns profits equal to

$$(5) \quad \pi_i = x_i(A - z + v(x_i^e))$$

when it has sales of  $x_i$  and total output is  $z$ .

When all  $n$  products are compatible,  $y_i^e = \sum_{j=1}^n x_j^e \equiv z^e$ , for all  $i$ . Therefore, when total output is  $z$  and firm  $i$  has sales of  $x_i$ , the firm's gross profits are

$$(6) \quad \pi_i = x_i(A - z + v(z^e)),$$

from which we must subtract  $F_i$  to get profits net of the fixed costs of compatibility.

### C. Fulfilled Expectations Equilibrium

Our equilibrium concept is that of fulfilled expectations Cournot equilibrium, where each firm chooses its output level under the assumptions that: (a) consumers' expecta-

<sup>3</sup>It is for this reason that negative values of  $r$  make sense; the redefined variable  $r$  measures the excess of a consumer's basic willingness to pay over the marginal production costs of an additional unit.

tions about the sizes of the networks,  $(y_1^e, y_2^e, \dots, y_n^e)$ , are given; and (b) the actual output level of the other firms,  $\sum_{j \neq i} x_j \equiv x_{-i}$ , is fixed.

Assumption (b) is the standard Cournot assumption. For any fixed set of consumer expectations, the problem is equivalent to the standard linear demand Cournot model with constant marginal costs. Assumption (a) is relaxed in the Appendix, which considers the case in which the  $y_i^e$ s are formed *after* the firms have selected their output levels. Differentiating equation (5) and rearranging terms, the first-order conditions  $d\pi_i/dx_i = 0$  imply that the equilibrium sales levels  $(x_1^*, x_2^*, \dots, x_n^*)$  must satisfy

$$(7) \quad x_i^* = A + v(y_i^e) - \sum_{j=1}^n x_j^* \quad \text{for } i=1, 2, \dots, n.$$

Note that the right-hand side of equation (7) equals  $p_i$ .

For any given set of expectations, we can solve equation (7) simultaneously for the  $x_i^*$ s to obtain the unique Cournot equilibrium that corresponds to that set of expectations:

$$(8) \quad x_i^* = \left\{ A + nv(y_i^e) - \sum_{j \neq i} v(y_j^e) \right\} / (n+1) \quad \text{for } i=1, 2, \dots, n.$$

The outcome is just the standard linear demand Cournot equilibrium where the differences in  $v(y_i^e)$  are analogous to production cost differences. Equation (8) defines a function that maps expectations  $(y_1^e, y_2^e, \dots, y_n^e)$  into Cournot equilibrium network sizes  $(y_1^*, y_2^*, \dots, y_n^*)$  for a given pattern of compatibility. Let  $\Gamma(y^e)$  denote this function.

In the absence of rationality constraints on consumer expectations, there is a Cournot equilibrium for any set of expectations. But for most sets of expectations, the expectations will not be fulfilled in the corresponding Cournot equilibrium; the actual network

sizes are not equal to the expected ones. Although it is possible that (in the short run, at least) consumers could be mistaken about network sizes, it is useful to limit the set of possible equilibria by imposing the restriction that expected sales be equal to actual sales in equilibrium. Formally, our equilibrium notion is that of Fulfilled Expectations Cournot Equilibrium (*FECE*), where a *FECE* is an  $n$ -vector of network sizes  $y^* = (y_1^*, y_2^*, \dots, y_n^*)$ , such that  $y^* = \Gamma(y^*)$ . If consumers expect the network sizes to be  $y^*$ , then in the corresponding Cournot equilibrium the network sizes will indeed equal  $y^*$ ; consumers' expectations will be fulfilled.

#### D. Welfare Formulae

Given the cost and demand assumptions that we have made, profits and welfare can be written as functions of the firms' individual levels of output. By equation (7), in equilibrium, firm  $i$ 's output level is equal to the price that firm receives. Thus, the  $i$ th firm's profits in equilibrium are  $\pi_i = (x_i^*)^2$ . We will denote aggregate profits by  $\pi \equiv \pi_1 + \dots + \pi_n$ .

The surplus that a consumer derives from joining a network depends on the *actual* size of the network; in equilibrium, the actual size will equal that network's *expected* size. By equations (1) and (3), when market output is  $z$ , a type  $r$  consumer expects to derive surplus of  $r + z - A$  from joining a network. Only those consumers for whom  $r$  is greater than  $A - z$  join a network; the other consumers stay out of the market and derive no surplus. Integrating over all consumers who do enter the market, we obtain consumers' expected surplus

$$(9) \quad S(z) = \int_{A-z}^A (p + z - A) dp = z^2/2.$$

In any fulfilled expectations equilibrium, expected and actual consumers' surplus will be equal, and we can use equation (9) when discussing actual consumers' surplus.

We take the sum of producers' and consumers' surplus as our social welfare measure. Hence, in any fulfilled expectations

Cournot equilibrium, welfare (gross of the fixed costs of compatibility) is given by

$$(10) \quad W(x_1, \dots, x_n) = \pi(x_1, \dots, x_n) + S(x_1 + \dots + x_n) = \sum_{i=1}^n x_i^2 + z^2/2.$$

## II. The Characterization of Equilibria

In this section, we examine the structure of fulfilled expectations equilibria for compatible and incompatible products, respectively.

### A. Complete Compatibility

Suppose that any two products are compatible with one another. Then there is a single network of expected size  $z^e = \sum_{i=1}^n x_i^e$ , and for all  $i$ ,  $y_i^e = z^e$ . Equation (8) becomes

$$(11) \quad x_i^* = (A + v(z^e))/(n+1) \quad \text{for } i=1, 2, \dots, n.$$

If we impose the fulfilled expectations requirement that  $z^e = x_1^* + \dots + x_n^*$  and sum equation (11) over all  $i$ , we obtain

$$(12) \quad z^c = (n/(n+1))(A + v(z^c)),$$

where  $z^c$  denotes the fulfilled expectations equilibrium value of total output when the products are compatible. Under our assumptions on  $v(\cdot)$ , equation (12) has a unique solution, as is clear from Figure 1. This unique compatible-products equilibrium is symmetric:  $x_i^c = z_i^c/n$  for all  $i$ . We have shown:

**PROPOSITION 1:** *When all products are mutually compatible, there is a unique FECE. It is symmetric, and the aggregate level of output is given implicitly by equation (12).*

As the number of firms becomes increasingly large, the compatibility equilibrium converges to the perfectly competitive equilibrium;  $z^c$  approaches  $A + v(z^c)$ , and the hedonic price,  $A + v(z^c) - z^c$ , approaches the marginal cost level of zero.

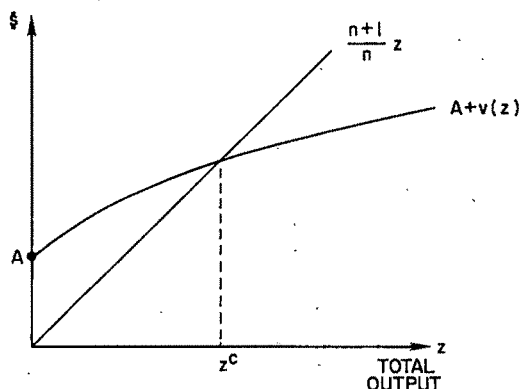


FIGURE 1. EQUILIBRIUM WITH COMPLETE COMPATIBILITY

### B. Complete Incompatibility

Now we consider the case where any two brands are incompatible with one another so that  $y_i^e = x_i^e$ . In equilibrium, each firm  $i$  is optimizing given the actions of the other firms,  $x_j$ ,  $j \neq i$ , and consumers' expectations,  $x_i^e$ . Using equation (7) in conjunction with the fulfilled expectations condition  $x_i = x_i^e$ , we have  $x_i = A + v(x_i) - z$ , or

$$(13) \quad \sum_{j \neq i} x_j = A + v(x_i) - 2x_i \quad \text{for } i=1, 2, \dots, n.$$

For a given value of  $x_{-i}$ , equation (13) can be solved for  $x_i$ . The graph of equation (13) is called firm  $i$ 's *equilibrium reaction correspondence*. One possible shape of this correspondence is illustrated in Figure 2.<sup>4</sup>

The equilibrium reaction correspondence should not be confused with a standard reaction function. The latter merely states firm  $i$ 's best response to the other firms, given consumer expectations. There will be a different reaction function for each set of expectations. The equilibrium reaction corre-

<sup>4</sup>It is also possible that firm  $i$ 's reaction schedule is strictly downward sloping. This will occur if and only if  $v'(0) < 2$ .

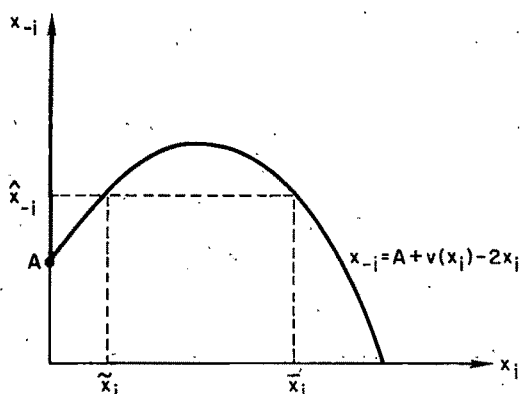


FIGURE 2. FIRM  $i$ 'S EQUILIBRIUM REACTION CORRESPONDENCE

spondence gives the set of points such that if the other firms played  $x_{-i}$  and consumers expected brand  $i$  to have a network size of  $x_i$ , then  $x_i$  would in fact be firm  $i$ 's best response. Suppose the other firms set their output at  $\hat{x}_{-i}$  in Figure 2. Then firm  $i$ 's best response to  $\hat{x}_{-i}$  would fulfill consumer expectations if these expectations were either  $\tilde{x}_i$  or  $\bar{x}_i$ . Note that firm  $i$  treats consumer expectations as exogenous, and thus the firm does not choose between  $\tilde{x}_i$  and  $\bar{x}_i$ .

In Figure 2, firm  $i$ 's equilibrium reaction correspondence is drawn to include the  $x_{-i}$  axis for  $x_{-i} > A$ . This part of firm  $i$ 's reaction schedule is not derived from equation (13), which only applies when  $x_i > 0$ . Instead it is derived from the corner condition,  $d\pi_i/dx_i < 0$  at  $x_i = 0$ . When  $x_i^e = x_i = 0$ ,  $d\pi_i/dx_i = A - x_{-i}$ , so it is optimal for firm  $i$  to set  $x_i = 0$  if  $x_{-i}^e = 0$  and  $x_{-i} > A$ .

Having derived the firms' fulfilled expectations reaction schedules, we turn now to the characterization of equilibria. There are three types of equilibria that are possible when the networks of competing firms are incompatible: (i) symmetric oligopoly with  $n$  active firms; (ii) symmetric oligopoly with  $k < n$  active firms, which we call natural oligopoly; and (iii) asymmetric oligopoly.

#### 1. Symmetric Oligopoly.

**PROPOSITION 2:** When each brand is incompatible with all  $(n-1)$  of the other brands,

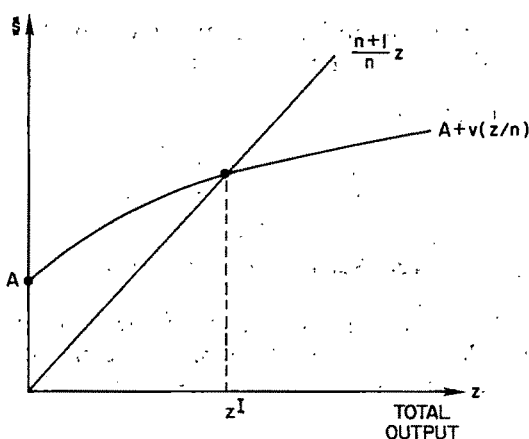


FIGURE 3. UNIQUE SYMMETRIC EQUILIBRIUM WITH COMPLETE INCOMPATIBILITY

there exists a unique symmetric equilibrium in which  $x_i = z^I/n$  and aggregate sales,  $z^I$ , are given implicitly by

$$(14) \quad ((n+1)/n)z^I = A + v(z^I/n).$$

#### PROOF:

Taking  $x_i = z/n$  and adding up equation (13) for  $i=1, \dots, n$ , gives  $(n-1)z = nA + nv(z/n) - 2z$ . Rearranging yields equation (14), which has a unique solution, as Figure 3 illustrates.

2. *Natural Oligopoly (Not all Firms Active).* While a unique symmetric equilibrium always exists, there are asymmetric equilibria that exist for certain parameter values. Given the symmetry in the equilibrium response correspondences, such asymmetric equilibria always come in sets (where the elements of the set differ from each other only by the transposition of the firms' indices). One such type of equilibrium entails some firms exiting the market (i.e., producing no output) and the other firms behaving as oligopolists with a diminished number of competitors.

**PROPOSITION 3:** A symmetric equilibrium with  $k$  active firms exists if and only if  $v(A/k) \geq A/k$ .



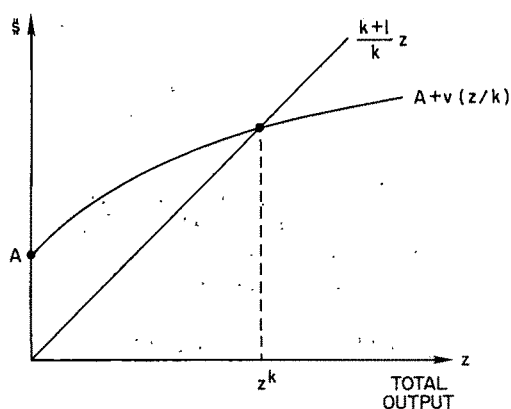


FIGURE 4. NATURAL OLIGOPOLY

## PROOF:

Suppose that  $k$  firms each produce  $x_i = z/k$  units of output, and the remaining  $n - k$  firms produce no output. Adding up equation (13) for the  $k$  active firms, we obtain  $(k - 1)z = kA + kv(z/k) - 2z$ , or

$$(15) \quad ((k + 1)/k)z = A + v(z/k).$$

As Figure 4 illustrates, there will be a unique solution to equation (15). Let  $z^k$  denote this solution.

We must check that the remaining  $n - k$  firms do not have incentives to produce positive output, that is, that  $A + v(0) - z^k = A - z^k < 0$ . Again from Figure 4, it is clear that  $z^k \geq A$  iff  $A + v(A/k) \geq ((k + 1)/k)A$ , or  $v(A/k) \geq A/k$ .

**COROLLARY 3.1:** *For any  $k \leq n - 1$ , if a  $k$ -active-firm symmetric equilibrium exists, then a  $(k + 1)$ -active-firm symmetric equilibrium exists.*

**COROLLARY 3.2:** *For any  $k \leq n - 1$ , if a  $k$ -active-firm symmetric equilibrium exists, then  $z^k < z^{k+1}$ .*

Both corollaries follow from the concavity of  $v(\cdot)$ . Note that equilibrium with  $k = 1$  (the monopoly outcome) or some other low value of  $k$  is more likely to obtain when consumers' basic willingness to pay for the good is low (so that  $A$  is low) or when the

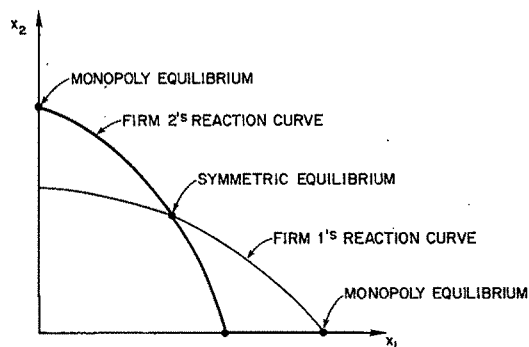


FIGURE 5. NATURAL MONOPOLY EQUILIBRIA

network effects are strong (so that  $v(A)$  is large for a given  $A$ ).

Proposition 3 shows that the network externalities are similar to fixed costs in that they can lead to a limited number of active producers. The analogy between fixed costs and network externalities is not complete, however. This point is demonstrated by Corollary 3.1 and the fact that for the given set of demand conditions, an  $n$ -active-firm equilibrium exists for arbitrarily large  $n$ . In the case of fixed costs, one cannot squeeze an arbitrarily large number of active producers into the industry.

Figure 5 shows the reaction curves for a case in which a natural monopoly equilibrium exists. It is interesting to note that the monopolist's profits, may be *lower* than the profits of a duopolist in the 2-active-firms symmetric equilibrium. In other words, a monopoly may benefit from entry. This unusual result follows from the fulfilled expectations condition: a monopolist will exploit his position with high prices and consumers know this. Thus, consumers expect a smaller network and are willing to pay less for the good. If the monopolist could commit himself to higher sales he would be better off, but this commitment is not credible so long as he is the sole producer.<sup>5</sup>

**3. Asymmetric Oligopoly.** The third possible equilibrium configuration is one in which

<sup>5</sup>We consider the case where commitment is feasible in the Appendix.

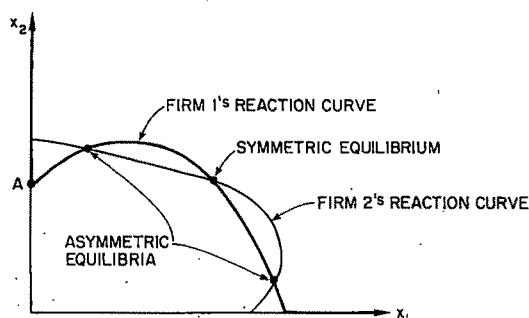


FIGURE 6. ASYMMETRIC DUOPOLY

$k \geq 2$  firms produce positive but unequal levels of output. We have constructed examples of such equilibria, although they are difficult to characterize in general. These asymmetric equilibria verify the intuition that a firm may be successful and enjoy a large market share simply because it is expected to by consumers.

One example of an asymmetric duopoly equilibrium is shown qualitatively in Figure 6. Despite a linear demand curve and a concave network valuation function  $v(y)$ , a variety of possibilities may arise. Figure 6 shows a situation in which asymmetric equilibria exist as well as symmetric and natural monopoly equilibria. In other cases, the only stable equilibria are asymmetric ones.

### C. Partial Compatibility

When there are more than two firms, the extent of product compatibility may fall in between complete incompatibility and industrywide compatibility. Assuming that the compatibility relation is symmetric and transitive, the pattern of compatibility can be characterized by the set of compatibility groups,  $G^j$   $j=1, \dots, J$ , where all of the brands within a given group are mutually compatible with each other and are incompatible with any nonmember brands.<sup>6</sup> Thus,

<sup>6</sup>The  $G^j$ ,  $j=1, \dots, J$ , form a partition of the set  $\{1, 2, \dots, n\}$ .

if firm  $i$  is in group  $G^j$ ,

$$y_i = \sum_{k \in G^j} x_k \equiv y^j.$$

For a firm  $i$  in group  $j$ , the first-order condition is  $x_i = A - z + v(y^j)$ . Thus, all firms in a given group will choose the same level of output,  $x^j$ . Let  $m^j$  denote the number of firms in compatibility group  $j$ . Then, in equilibrium, for all  $x^j > 0$  we must have

$$(16) \quad x^j = A - z + v(m^j x^j).$$

Equation (16) has the same qualitative properties as our earlier equilibrium conditions, and similar types of equilibria will arise. Here, we will not characterize these equilibria directly. In the next subsection, however, we will compare the equilibria that obtain under different degrees of compatibility.

### D. The Output Effects of Compatibility Changes

In analyzing compatibility, it is important to understand the effects of an increase in compatibility on the equilibrium levels of output. What happens to output levels if two compatibility groups "merge" to form a new group where all of the brands in the post-merger group are compatible with one another?

**PROPOSITION 4:** *The level of total output is greater under industrywide compatibility than in any equilibrium with less than complete compatibility.*

#### PROOF:

For all firms with positive levels of output,  $x_i = A + v(y_i) - z$ . Adding up over all firms and rearranging gives us  $(n+1)z = nA + \sum v(y_i)$ , as illustrated in Figure 7. Under complete compatibility,  $y_i = z$  for all firms. Absent complete compatibility,  $y_i < z$  for at least one firm. Thus, the curve  $nA + nv(z)$  lies above  $nA + \sum v(y_i)$  where the  $y_i$ 's are

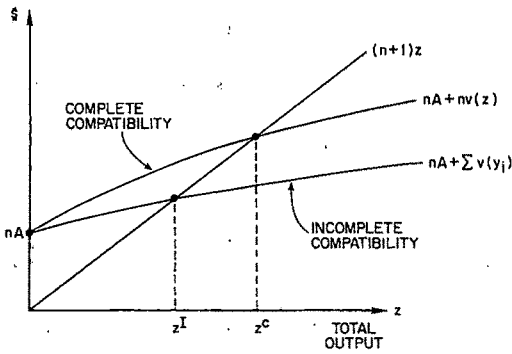


FIGURE 7. COMPLETE VS. INCOMPLETE COMPATIBILITY

determined under incomplete compatibility. Referring to Figure 7, we see that the equilibrium level of  $z$  is greater under industry-wide compatibility.

When the move to increased compatibility does not result in *complete* compatibility, total output need not rise. The following proposition, however, states a sufficient condition for industry output to increase.

**PROPOSITION 5:** *Suppose that two groups of firms make their products mutually compatible. If premerger total output is less than  $A$ , then in any postmerger equilibrium: (a) the average output of the firms in the merging coalitions will rise; (b) the output of any firm not in the merging coalitions will fall; and (c) industry output will rise.*

**PROOF:**

Let  $\hat{x}^j$  denote the premerger output level of a firm in coalition  $j$  and  $\hat{z}$  denote premerger total output. By equation (16),  $\hat{x}^j = A - \hat{z} + v(m^j \hat{x}^j)$ . Figure 8 illustrates this condition, where we have made use of the fact that industry output is less than  $A$ .

Label the compatibility groups that merge as 1 and 2. Let  $\tilde{x}^j$  and  $\tilde{z}$  denote the postmerger output levels analogous to the  $\hat{x}^j$  and  $\hat{z}$ . If total output falls, then for  $j \geq 3$ ,  $A - \tilde{z} + v(m^j \tilde{x}^j)$  will lie everywhere above  $A - \hat{z} + v(m^j \hat{x}^j)$ , and from Figure 8 we see that  $x^j$  will increase (i.e.,  $\tilde{x}^j > \hat{x}^j$ ). By similar argu-

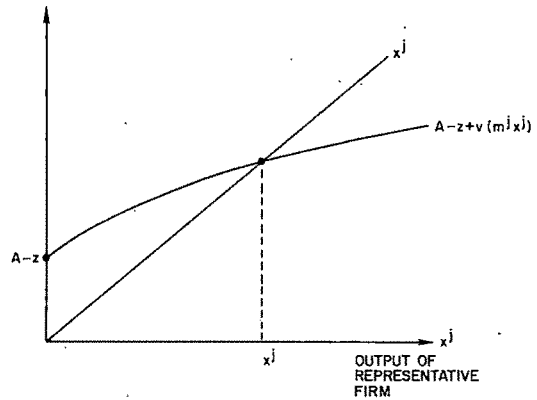


FIGURE 8. EQUILIBRIUM OUTPUT OF A FIRM IN COALITION  $j$

ments,  $\tilde{z} = \hat{z}$  implies  $\tilde{x}^j = \hat{x}^j$  for  $j \geq 3$  and  $\tilde{z} > \hat{z}$  implies  $\tilde{x}^j < \hat{x}^j$  for  $j \geq 3$ .

Now, consider the firms in the merged coalitions. For these firms (i.e.,  $j=1,2$ ),  $A - z + v(m^j x^j) < A - \tilde{z} + v(m^1 \tilde{x}^1 + m^2 \tilde{x}^2)$  as long as  $x^1$  and  $x^2$  are positive. The effect of compatibility is to shift up the  $A - z + v$  curve when viewed as a function of  $x^j$  (as in Figure 8). The effect of  $z$  on this curve is the same as for  $j \geq 3$ . Therefore if  $\tilde{z} \leq \hat{z}$ , then  $\tilde{x}^j > \hat{x}^j$  for  $j=1,2$ .

Suppose that  $\tilde{z} \leq \hat{z}$ . Then  $\tilde{x}^j \geq \hat{x}^j$  for all  $j$ , with strict inequality for  $j=1,2$ . But  $z = \sum m^j x^j$ , so we have a contradiction. Therefore, industry output rises due to the increased compatibility;  $\tilde{z} > \hat{z}$ . We already have shown that an increase in industry output implies that firms not in the merging coalitions produce less;  $\tilde{z} > \hat{z}$  implies  $\tilde{x}^j < \hat{x}^j$  for  $j \geq 3$ . Thus, it must be the case that the firms in the merging coalitions produce more:  $m^1 \tilde{x}^1 + m^2 \tilde{x}^2 > m^1 \hat{x}^1 + m^2 \hat{x}^2$ .

Note that Proposition 5 does *not* state that the new level of per firm output for the enlarged coalition will be larger than both  $\hat{x}^1$  and  $\hat{x}^2$ ; it states only that the new level of output will be larger than their mean. Note also that by Proposition 4, if total output under complete compatibility,  $z^c$ , is less than  $A$ , then  $z < A$  in *any* equilibrium. From equation (12),  $z^c < A$  if  $v(A) < A/n$ . So,

$v(A) < A/n$  is sufficient for  $\hat{z} < A$ , the hypothesis of Proposition 5.

When industry output in the premerger equilibrium equals or exceeds  $A$ , an increase in compatibility may be accompanied by a reduction in both industry output and the average level of output of the firms in the merging coalitions. To see this point, consider the following example. There are  $n$  firms,  $v(A) \geq A$ , and initially there is complete incompatibility. By Proposition 2, there exists a unique symmetric equilibrium in which all  $n$  firms are active. By Proposition 3, there exist  $n$  natural monopoly equilibria, each of which entails the natural monopolist producing  $A$  or more units of output. The natural monopolist's output level is less than that of the industry in the  $n$ -active-firm symmetric equilibrium (Corollary 3.2).

Suppose that the industry initially is in the  $n$ -active-firm symmetric equilibrium, and that some, but not all, of the firms form a compatibility coalition. The natural monopoly equilibria where a nonmerging firm is the natural monopolist will remain equilibria after the merger. Thus, while the premerger equilibrium entailed all firms active, the postmerger equilibrium could be one in which all but one firm shuts down. As we have just stated, industry output falls in this case. Moreover, the firms who formed the compatibility coalition are among those who were active in the premerger equilibrium, but are inactive in the postmerger equilibrium.

### III. The Private and Social Incentives for Network Compatibility

To this point, we have treated the compatibility of the products as an exogenous characteristic of the market. In most markets where network externalities are important, the compatibility of the products will be the result of explicit decisions by the firms. When the network externalities are large, the choice of whether to make the products compatible will be one of the most important dimensions of market performance.

There are many cases in which firms will disagree on the desirability of making their products compatible; the move to compati-

bility may increase the profits of some firms while lowering the profits of others. Thus, we must be careful to specify the mechanism by which compatibility is achieved and whether side payments among firms are feasible.

It is useful to think of there being two basic technologies by which compatibility can be achieved. First, compatibility may arise through the *joint adoption of a product standard*, where a given set of firms must act together to make their products compatible with one another. The CPM operating system for personal computers and the broadcast television standards are two examples. Second, compatibility can be achieved through the *construction of an adapter*, where a single firm can act unilaterally to make its product compatible with those of another firm or group of firms. In the 1960's, for example, Honeywell developed a program that would allow its mainframe computers to run programs initially written for IBM hardware.<sup>7</sup> At present, there is intense competition, both in the market and in the courts, as the video game manufacturers such as Coleco develop adapters that allow their hardware to run the video game programs of competing firms. In other cases there need not be a physical adapter; one firm can adopt another network's specifications for its product design.

When the firms cannot make side payments among one another and when the compatibility mechanism is a product standard, the products of a given set of firms will be made compatible if and only if *all* of these firms would earn greater profits as a result. In contrast, when the compatibility mechanism is an adapter, and side payments are infeasible, the products of two firms will be made compatible if *either* firm would find the move to be profitable.

When side payments are feasible, the firms will make their products compatible if and only if the change in the profits of firms within the set that can make side payments to one another exceeds the joint costs of compatibility.

<sup>7</sup>See Brock, p. 78.

We will examine several cases that vary in terms of the compatibility technology and feasibility of side payments. In analyzing the private incentives for compatibility, we will look at each firm's change in profits,  $\Delta\pi_i = \pi_i^c - \pi_i^I$ , and the change in their joint profits  $\Delta\pi = \sum_{i=1}^n \Delta\pi_i$ , and we compare these with the costs of compatibility. For most of this section we will take these costs to be purely fixed costs. The social incentives for compatibility are given by  $\Delta W = W^c - W^I$ . We denote the change in consumers' surplus by  $\Delta S = S^c - S^I$ .

#### A. The Incentive when Side Payments among All Firms are Feasible

If side payments among all firms are feasible, then a set of side payments could be constructed so that all firms' profits would be increased individually if and only if compatibility would raise joint profits; the private incentives are given by the change in industrywide profits under either compatibility technology.<sup>8</sup> The change in social welfare,  $\Delta W = \Delta\pi + \Delta S$ ; so the social and private incentives will diverge when the move to compatibility changes the level of consumers' surplus. Since  $S(z) = z^2/2$ , consumers' surplus will go up if and only if output does. By Proposition 4, we know that output and, hence, consumers' surplus will rise with the move to full compatibility. Thus, if  $\Delta\pi > 0$ , then  $\Delta W = \Delta\pi + \Delta S > 0$ .

**PROPOSITION 6:** *When compatibility costs are purely fixed costs, any move to complete compatibility that raises industry profits is socially beneficial.*

Proposition 6 states that firms' incentives for compatibility are not socially excessive. In fact, they may be inadequate. Since  $\Delta S > 0$ ,  $\Delta W > \Delta\pi$ . If the industrywide costs of

compatibility,  $F$ , satisfy  $\Delta\pi < F < \Delta W$ , then the private firms will fail to adopt a socially desirable system of compatibility. The reason is that the firms cannot appropriate all the benefits of compatibility.<sup>9</sup>

**PROPOSITION 7:** *Even when arbitrary side payments among all firms are feasible, profit-maximizing firms may fail to achieve complete compatibility in cases where complete compatibility is socially optimal.*

A similar pair of results can be derived from Proposition 5 for markets in which the initial equilibrium entails  $z < A$  and compatibility is increased (although not necessarily to industrywide compatibility). This result too relies on the fact that the move to increased compatibility raises total output under the stated conditions.

#### B. The Adoption of an Industry Standard

When the compatibility mechanism is the adoption of an industry standard, the firms must jointly decide to make their networks compatible. Any firm can veto the move to compatibility. Therefore, if no side payments are feasible, the standard will be adopted if and only if all firms joining the standard benefit from its creation. If we assume that firm  $i$  bears a cost  $F_i$  to adopt the standard and that side payments are infeasible, then adoption will occur if and only if  $\Delta\pi_i > F_i$  for all adopters of the standard.

Suppose that side payments are feasible only when made among firms achieving compatibility. Such side payments might take the form of licensing fees or compensation for the expenses of making the products compatible, for example. In this case, a sufficient condition for achieving compatibility with a standards technology is that the joint profits of the firms achieving compatibility rise.

<sup>8</sup>Instead of payments made in return for not achieving compatibility, we might see expenditures on legal proceedings aimed at blocking the move to compatibility.

<sup>9</sup>This result is analogous to the fact that a monopolist may be unable to earn a positive profit by providing a socially useful product if there are fixed costs and he or she cannot perfectly price discriminate.

It is clear that, when the compatibility technology is a standard, allowing cost sharing will raise the likelihood of the firms choosing compatibility. If each firm prefers adoption of the standard ( $\Delta\pi_i > F_i$  for all  $i$ ), then the firms in aggregate will ( $\sum \Delta\pi_i > \sum F_i$ ), while the converse is not true. Thus, we can strengthen Proposition 6:

**PROPOSITION 8:** *The private standardization rule is more stringent when cost sharing is infeasible than when it is feasible. The set of cases in which the firms fail to adopt a socially beneficial standard is therefore larger. It remains true that any privately profitable industrywide standard is desirable.*

To see the effect of cost sharing, suppose that there are only two firms in the industry. The compatibility equilibrium is symmetric, so the firm with the smaller incentive to adopt a standard is the one with initially higher profits, that is, the initially larger producer, say firm 1. If the initial equilibrium is symmetric and  $F_1 = F_2 = F$ , then  $\Delta\pi_1 - F = \Delta\pi_2 - F$  and the presence or absence of cost sharing or other side payments is irrelevant. If the initial equilibrium is asymmetric, however, the condition for standardization,  $\Delta\pi_1 > F$  is *strictly* more stringent than the adoption condition with side payments,  $\Delta\pi > 2F$ . The problem is that the larger firm will lose market share to its smaller rival as a result of standardization. If it can unilaterally block standardization, it may do so, despite the fact that its rival and consumers would benefit. Permitting cost sharing and other side payments will help alleviate the problem of insufficient private adoption incentives for moves to complete compatibility.

When the compatibility increase is to less than complete compatibility, private incentives may be excessive, and such cost sharing may exacerbate the problem. There are two reasons why private incentives may be excessive. First, when the increase is to less-than-complete compatibility, total output and consumers' surplus may fall, so that  $\Delta\pi > \Delta W$ . Second, there will be some firms that are not members of the groups making their products compatible. As shown in the proof

of Proposition 5, these firms may produce less output and thus have lower profits in the new equilibrium. Absent side payments from these firms, the firms considering compatibility will not take the losses of other firms ( $\sum_{j \neq i} \Delta\pi_j < 0$ ) into account. The social incentives, however, depend on the profits of all firms;  $\Delta W = \Delta\pi_i + \sum_{j \neq i} \Delta\pi_j + \Delta S$ .

**PROPOSITION 9:** *When the increase in compatibility leads to less-than-industrywide compatibility, the private incentives to standardize may be excessive.*

### C. The Construction of an Adapter

In the adapter case, a firm unilaterally can act to make its product compatible with those of another network. In contrast with the adoption of an industry standard, if side payments to block compatibility are not feasible, then the adapter will be constructed as long as at least one firm earns increased profits from compatibility. When the compatibility mechanism is an adapter, the most reasonable assumption about the costs of compatibility is that the firm that constructs the adapter is the only one to bear the cost,  $F$ . Thus, firm  $i$ 's private incentive to construct an adapter is  $\Delta\pi_i - F$ , while the social incentive is  $\Delta\pi_i + \sum_{j \neq i} \Delta\pi_j + \Delta S - F$ . The difference,  $\Delta\pi_{-i} + \Delta S$ , may in general be either positive or negative, implying that firm  $i$ 's incentives to construct an adapter may be too low or too high from a social welfare point of view.

To see how the private and social incentives differ, suppose there are only two firms. Since the smaller firm in the initial equilibrium has the most to gain by the move to a symmetric equilibrium with compatibility, we need look only at the incentives of a firm with an initial market share of not more than 50 percent—firm 2, say. Since  $\Delta\pi_2 > \Delta\pi_1$ , the private decision will be to become compatible if and only if  $\Delta\pi_2 > F$ .<sup>10</sup> Compatibility is

<sup>10</sup> We are ignoring the possibility that the firms play a waiting game in which each hopes (in vain) that the

socially optimal if and only if  $\Delta W > F$ . The divergence between the social and private incentives is given by  $\Delta W - \Delta\pi_2 = \Delta\pi_1 + \Delta S$ . By Proposition 4, we know that  $\Delta S > 0$ ; the fact that consumers enjoy some of the benefits of compatibility tends to make the private incentives too low. On the other hand, it is not true in general that  $\Delta\pi_1 > 0$ , so we cannot conclude in general that the private adoption decision is too conservative.

One case in which the private adoption incentives are too low is when the initial equilibrium is symmetric. In that case,  $\Delta\pi_1 = \Delta\pi_2$ , so that if  $\Delta\pi_i > 0$  for one firm, then the change in profits is positive for the other firm as well. As a result,  $\Delta W > \Delta\pi_i$  whenever  $\Delta\pi_i > 0$ , and we have

**PROPOSITION 10:** *Suppose there are only two firms (or coalitions). If the incompatibility equilibrium is symmetric and there are no side payments, then the private incentives to construct an adapter are too low.*

When there are only two coalitions, permitting side payments to share the costs of the adapter would promote efficiency when the incompatibility equilibrium is symmetric. The adapting coalition confers benefits both on its rival, which free rides on compatibility, and on consumers. Side payments can help solve the free-riding problem, but (by Proposition 6) still leave the firms with insufficient incentives.

When the incompatibility equilibrium is asymmetric, it is possible that the initially larger firm (which does not build the adapter) loses so much market share when the smaller firm builds an adapter that its profits fall, that is,  $\Delta\pi_1 < 0$ . If this effect dominates the increase in consumers' surplus, that is, if  $\Delta\pi_1 + \Delta S < 0$ , then  $\Delta W < \Delta\pi_2$  and firm 2's compatibility incentives are excessive. The increase in firm 2's market share at the expense of firm 1 is a private gain for which there is no corresponding social benefit.

**PROPOSITION 11:** *Suppose there are only two coalitions. A coalition with an incompatibility market share of less than 50 percent may have socially excessive incentives to construct an adapter.*

This result is most likely to obtain when the incompatibility equilibrium entails one coalition having a very small market share, as in the monopoly equilibrium.

When the means of achieving compatibility is the construction of an adapter, one firm may attempt to make the networks compatible even though the other firms would prefer that the networks remain incompatible. In such cases, the latter firms may be willing to make expenditures to block compatibility, perhaps through legal channels (currently, there are numerous court cases involving video game and personal computer compatibility). It is not possible to say in general whether such expenditures promote or diminish efficiency. In some cases,  $\Delta\pi_1 < 0$ ,  $\Delta\pi_2 > F$ , and  $\Delta W < F$  as noted in Proposition 11, and firm 1's ability to block the adapter will raise efficiency (if the blocking costs themselves are not too high). In other cases,  $\Delta\pi_1 < 0$ ,  $\Delta\pi_2 > F$ , and  $\Delta W > F$ , so blocking the adapter would reduce social welfare, even if the blocking costs are zero. For a given cost and demand structure, one can determine whether  $\Delta W$  exceeds  $F$  or not, but it is not possible to determine this relationship simply on the basis of  $\Delta\pi_2$  exceeding  $F$ .

#### D. Extensions and Generalizations

We have made some restrictive assumptions in order to simplify our analysis of the incentives for network compatibility. It is useful to put the results that we have obtained into perspective by discussing the general nature of the divergence between the social and private incentives to achieve compatibility. Essentially, there are two sources of distortion. In making its compatibility decision, each firm ignores the effects that this move will have on: 1) the level of consumers' surplus; and 2) the profits of the other firms.

---

other will build an adapter even though each would privately benefit from building the adapter itself.

Consider the first effect. When the move to compatibility raises consumers' surplus, the firms' incentives tend to be too low. Conversely, when the move to compatibility lowers consumers' surplus, the firms are biased towards compatibility. The change in consumers' surplus can itself be decomposed into two components: (a) the change due to the shift in the level of total output; and (b) the change that arises when the marginal consumer values the network externality differently than does the average consumer.

(a) The level of consumers' surplus is an increasing function of the level of total output. When the only costs of compatibility are fixed, we showed that the move to complete compatibility raises output and, hence, consumers' surplus. In this case,  $\Delta\pi$  is less than  $\Delta W$ . Once we relax the assumption that the move to compatibility has no impact on marginal costs, however, output may be lower with complete compatibility than without. The adoption of an industry standard or the construction of an adapter will necessitate the redesign of some or all of the products, which may lead to shifts in the variable costs of production (either upwards or downwards). Whereas the fixed costs of compatibility do not affect the equilibrium output level, changes in marginal costs do. In particular, when the increase in marginal costs is sufficiently large relative to the network effects, total output will be lower under complete compatibility than under incompatibility. In these cases, consumers' surplus will fall as a result of the move to complete compatibility and  $\Delta\pi$  is greater than  $\Delta W$ —the firms' joint incentive are excessive.<sup>11</sup>

<sup>11</sup>In fact, when compatibility raises producers' marginal costs, the firms may use the move to compatibility *solely* as a coordinating device to reduce their joint output (i.e., they may have incentives to make their products compatible even if there are no network externalities). This result is an example of the general theory of cost-based facilitating practices (see Steven Salop and David Scheffman, 1983; Katz and Harvey Rosen, 1985; and Jesus Seade, 1983): in an oligopoly, all firms may benefit from jointly increasing their costs because it induces them to reduce their collective output, which may raise their revenues by more than the increase in costs.

(b) The level of consumers' surplus also depends on the relationship between the marginal and inframarginal consumers' valuations of the good. For a given level of output, the firms must set prices low enough to attract the marginal consumer. The lower is the marginal consumer's valuation relative to the average consumer's valuation, the larger will be consumers' surplus.

In our model, all consumers value the network externality equally, and all consumers' valuations of the good rise equally when compatibility is achieved. Thus, for a fixed level of output, the firms can raise prices by just this amount and consumers' surplus is unaffected. More generally, consumers may differ in their valuations of the network externality. If the network externality is stronger for the marginal consumer, then the move to compatibility will raise his or her willingness to pay for the good by more than that of the average consumer. For a given level of output, the firms will be able to raise the price by more than the increase in the average consumer's willingness to pay for the product. Consumers' surplus will fall, and the joint private incentives will tend to be greater than the social incentives. Of course, if the network externality is smaller for the marginal consumer, then the bias will run in the other direction.<sup>12</sup>

The change in consumers' surplus puts a wedge between the change in joint profits and the change in total welfare. When side payments are not feasible, the decision to achieve compatibility depends on the individual profit levels of the firms, and there is a second wedge. The change in profits may be positive for some firms and negative for others. As we have shown, the relationship between the changes in firms' profits will depend on two factors. First, it will depend on their relative changes in market shares

<sup>12</sup>It is straightforward, but messy, to extend our model in this direction. There is nothing about this particular problem that is intrinsic to networks. For discussion of the general inability of prices to convey information about the preferences of inframarginal consumers, see A. Michael Spence (1975) and the references cited therein.



and revenues in moving to compatibility. When one group of firms gains market share and profits at the expense of another, the first group will be biased towards compatibility and the latter will be biased away from it. Second, the relationship will depend on the relative costs of compatibility that the producers incur. When the costs of compatibility fall more heavily on some firms than on others, there is a free-rider problem that tends to bias the firms away from compatibility.<sup>13</sup>

#### IV. Conclusion

We have developed a simple model to capture what we believe is a very significant element of competition in several important markets. Despite this simplicity, some general points emerge. First, the structure of the equilibria in our model confirms the importance of consumers' expectations in markets where network externalities are present. We have subjected expectations to a rationality constraint, but the expectations formation process remains an important element of the market to model explicitly. Given the possibility of multiple equilibria when products are incompatible, firms' reputations may play a major role in determining which equilibrium actually obtains. For example, the existence of a strong reputation for being a market share leader may explain IBM's rapid rise to preeminence in the personal computer market. It would also be useful to consider firms' expenditures to influence consumers' expectations, such as precommitments to a given level of software.

Turning to the compatibility decisions, although we would not want to draw policy conclusions at such an early stage in the analysis, our model does point to areas in which public policy can have an important impact. We have shown that the private decision will depend crucially on the decision locus (whether firms can act unilaterally or if consensus is required) and on the feasibility

of side payments. Public policy can influence both of these features. Patent and copyright laws are a significant determinant of whether the compatibility technology is better modeled as the joint adoption of an industry standard (when patents are strictly and broadly enforced), or as the unilateral construction of an adapter (when they are loosely enforced or narrowly applied). From Proposition 1, we know that if the costs of adapting are negligible, and there are no other entry barriers, the market will be perfectly competitive.<sup>14</sup>

Allowing firms to make side payments also may influence the likelihood of compatibility being adopted—upwards when the technology is an industry standard, and either upwards or downwards when the compatibility technology is an adapter. The discussion in Section III, Part D, also points out the need for policymakers to scrutinize the form of the side payments or royalties. Per unit charges may have the effect of implicit cartels by inducing outputs contractions. Finally, public policy can affect the costs of compatibility. Antitrust exemptions that allow industry groups to get together may lower the costs of achieving compatibility and thus make it more likely.

The model here is only a beginning. Explicitly dynamic, multiperiod models are needed to shed additional light on the behavior of markets in which network externalities are important. We hope that this paper will encourage further research in the area of network competition and public policy towards compatibility.

#### APPENDIX

In the text, we have examined a model where a firm's announcement of its planned level of output has no effect on consumer expectations. This model can be viewed as

<sup>13</sup>Here we are assuming that when the compatibility technology is an adapter, the costs fall more heavily on the adapting firm.

<sup>14</sup>This outcome may not be the socially optimal one. Absent the ability to earn rents from its network size (through incompatibility), a firm may not have incentives to make the investments necessary to obtain the network. The issues are exactly analogous to those encountered in the analysis of optimal patent policy.

one in which the firms are unable to commit themselves, so that only the output levels of the *FECE* are credible announcements. In this Appendix, we consider the opposite polar case in which firms *can* commit to announced output levels before consumers make their purchase decisions. Firm *i* commits itself to output level  $x_i$  and consumers make their purchase decisions by looking at  $v(y_i) - p_i$  across all brands. Firm *i* chooses its level of output taking the output level of the other firms as given. Thus, we have a standard Cournot equilibrium with demand-side economies of scale.

Given total output  $z$  and firm output  $x_i$ , firm *i* has profits of

$$x_i \{ A + v(y_i) - z \}.$$

Differentiating with respect to  $x_i$ , the first order conditions are

$$(A1) \quad A + v(y_i) - 2x_i - \sum_{j \neq i} x_j + x_i v'(y_i) = 0 \quad \text{for } i = 1, 2, \dots, n.$$

The only difference between equation (A1) and our earlier first-order condition is the addition of the term  $x_i v'(y_i)$ . This term captures the fact that firm *i* can directly influence consumers' expectations regarding its network size.  $x_i v'(y_i)$  is positive, so that firm *i*'s reaction curve will shift upwards in comparison with the earlier equilibrium reaction correspondence.

The analysis is essentially unchanged from that in the text; we simply substitute  $v(y_i) + x_i v'(y_i)$  for  $v(y_i)$  in firm *i*'s reaction function. Some additional assumptions on  $v$  are necessary to ensure that this substitute function is itself concave. In the case of complete incompatibility ( $y_i = x_i$  for all *i*), and a constant elasticity network externality function

$v(x) = \beta x^\alpha$ , we simply replace  $v(x) = \beta x^\alpha$  by  $v(x) + xv'(x) = \gamma x^\alpha$ , where  $\gamma = (1 + \alpha)\beta$ .

Qualitatively, the "commitment" equilibria differ from those analyzed in the text in the following ways. 1) Each firm's reaction correspondence becomes a reaction function, since it can "choose"  $x_i^e$  as well as  $x_i$  (see the discussion of Figure 2 in Section II, Part B). 2) Equilibrium entails greater output, as each firm accounts for the  $x_i v'(y_i)$  term, which shifts its reaction curve outwards. 3) No longer can a firm make greater profits in the  $k + 1$ -active-firm equilibrium than in the  $k$ -firm equilibrium.

## REFERENCES

- Brock, Gerald, "Competition, Standards, and Self-Regulation in the Computer Industry," in R. E. Caves and M. J. Roberts, eds., *Regulating the Product: Quality and Variety*, Cambridge: Ballinger, 1975.
- Katz, Michael L., and Rosen, Harvey S., "Tax Analysis in an Oligopoly Model," *Public Finance Quarterly*, January 1985, 13, 3-20.
- Kurdle, Robert T., "Regulation and Self-Regulation in the Farm Machinery Industry," in R. E. Caves and M. J. Roberts, eds., *Regulating the Price: Quality and Variety*, Cambridge: Ballinger, 1975.
- Oren, Shmuel S. and Smith, Stephen A., "Critical Mass and Tariff Structure in Electronic Communications Markets," *Bell Journal of Economics*, Autumn 1981, 12, 467-86.
- Salop, Steven and Scheffman, David, "Raising Rivals' Costs," *American Economic Review Proceedings*, May 1983, 73, 267-71.
- Seade, Jesus, "Prices, Profits and Taxes in Oligopoly," unpublished draft, University of Warwick, 1983.
- Spence, A. Michael, "Monopoly, Quality, and Regulation," *Bell Journal of Economics*, Autumn 1975, 6, 417-29.

# Asymmetric Information and Collusive Behavior in Auction Markets

By JONATHAN S. FEINSTEIN, MICHAEL K. BLOCK, AND FREDERICK C. NOLD\*

We present a theoretical and empirical analysis of the behavior of a bidder's cartel in a multiperiod auction market in which the purchaser is relatively uninformed. Our work establishes a tentative connection between the economics of information and collusion by concentrating on the cartel's informational monopoly, and its ability to both increase profits and mask its presence by passing misinformation to the purchaser.

In the models we develop, costs are assumed to be stochastic, and projects awarded in adjacent periods to be substitutes for one another. As a result, the purchaser will have incentive to intertemporally reallocate his demand, buying more or less of the current period's project, depending upon whether the current market price is lower or higher than the price which he expects to prevail in the future. A purchaser who needs to form such price expectations, yet is ignorant about production costs and market structure, will be particularly attracted to the auction mechanism, because the data contained in past auctions' bids can apparently be incorporated into an explicit structural model of expected future price. In fact, such a purchaser may hold auctions frequently merely to acquire information.

Such pleasant appearances are deceiving. Once the auction's bidders become aware of the informational use to which their bids are being put, they will have incentive to form a cartel, thereby extracting a premium for their

knowledge by misinforming the purchaser and skewing his intertemporal decision making to their advantage.

In some respects the mechanics of our model resemble the case of a multiproduct monopolist selling goods that are substitutes for one another. What is unique to our analysis is the dependence of one of the goods, future demand, on price expectations, which are an information resource. We are suggesting that information that is distributed asymmetrically among market participants is susceptible to the inefficiencies of collusion. In some ways the implications of our work are considerably stronger than this, because our cartel never reveals its true knowledge, and thus our market may be more inefficient than more standard monopolies, in which the good is always traded, albeit at noncompetitive prices. Furthermore, by explicitly modeling the purchaser's rational structural approach to price forecasting, we point out that complete rational expectations models may be more susceptible to exploitation by informational cartels than simpler forecasting techniques.

We have tested the implications of our model in a case study of North Carolina highway construction cartels. The highway construction industry is an auction market characterized by government procurement agencies who are uninformed relative to suppliers, and comes complete with a substantial record of known collusion in recent years. Thus it seems an ideal hunting ground for verification of our theoretical findings.

Our paper proceeds as follows. In Section I, we model a multiperiod competitive auction market, and describe a Bayesian mechanism through which the purchaser uses past data to improve the accuracy of his price expectations. Section II establishes theoretical results for the bidders' cartel which has incentive to form in this market: we describe

\*Massachusetts Institute of Technology, Cambridge, MA 02109 and University of Arizona, Tucson, AZ 85721, respectively. Nold was formerly of Rhodes Associates and is now deceased. We thank Richard Schmalensee and the participants of the Economic and Legal Organization Workshop at the University of Chicago for valuable suggestions. Part of this research was supported under research grant 81-IJ-CR0062 from the National Institute of Justice.

the specific strategies through which the cartel can misinform the purchaser, characterize the cartel's optimal short-run and long-run behavior, and mention extensions to our basic model. Section III presents our empirical results in highway construction, and some conclusions are offered in Section IV.

### I. The Auction Model

We begin the presentation of the formal model of our multiperiod auction market by summarizing its important features.

In a typical period  $t$ , a single purchaser solicits sealed bids that represent unit-cost offers on the period  $t$  project. After receiving the bids, the purchaser chooses the quantity, if any, of the period  $t$  project that he will purchase, and also decides whether or not to expend the fixed costs of holding an auction in period  $t + 1$ . Expected costs are identical across bidders and periods.<sup>1</sup> In addition, costs stochastically fluctuate, and as a result, the low bids received in different periods will fluctuate. We assume that projects in adjacent periods are partial substitutes for one another; hence, the purchaser will intertemporally substitute demand between periods in response to fluctuations in the low bid, and so his demand for period  $t$ 's project will depend not only on the period  $t$  low bid, but also on his expectation of period  $t + 1$ 's low bid. Since project costs are correlated, the purchaser can effectively use the information contained in earlier auctions' bids in forming his expectation of the period  $t + 1$  low bid.

To proceed, we define:

$v_t^i$  = the  $i$ th bidder's valuation of the unit cost of the period  $t$  project,

$Q_t$  = the quantity demanded by the purchaser in period  $t$ ,

$b_t^i$  = the  $i$ th bidder's bid in period  $t$ ,

$e_t$  = the purchaser's estimate of the low bid expected in period  $t$ , where  $e_t$  is formed

after the period  $t - 1$  bids have been submitted,

$P_t$  = the purchaser's decision whether or not to hold an auction in period  $t$ , with  $P_t$  made after the period  $t - 1$  bids have been submitted,

$R_t$  = the pool of available bidders in period  $t$ ,

$N_t$  = the number of actual bidders in period  $t$ ,

$m_t$  = the mean of the valuations  $v_t$  in period  $t$  where

$$m_t = \frac{1}{N_t} \sum_{i=1}^{N_t} v_t^i,$$

$s_t^2$  = the variance of the valuations  $v_t$  in period  $t$  where

$$s_t^2 = \frac{1}{N_t - 1} \sum_{i=1}^{N_t} (v_t^i - m_t)^2.$$

Note that our assumption that bidders will supply any quantity at their unit price bid is, in one context, equivalent to the specification of a constant marginal cost technology.<sup>2</sup>

The underlying uncertainty in our model arises because costs stochastically fluctuate around a fixed value. Specifically, bidders unit valuations  $v_t^i$  are independently drawn from a common distribution  $f$  (with cumulative distribution function  $F$ ) according to the rule:

$$(1) \quad v_t^i = \bar{p} + z_t^i \quad \text{where } \bar{p} \text{ is a constant,}$$

$$\text{and } z_t^i \sim N(0, \sigma_p^2).$$

The mean  $\bar{p}$  is identical across periods and bidders, relating cost estimates of projects awarded in different periods. The  $z_t^i$  represent idiosyncratic fluctuations in bidders' unit costs. Such fluctuations are assumed to result both from the transitory cost advantages and disadvantages of specific bidders, and from the specific attributes of the period  $t$  project.

<sup>1</sup>Our empirical work deals with highway construction, where each construction project has many line item components. Projects will often have common line items, correlating their overall values.

<sup>2</sup>If the bidder knows the purchaser's demand function, the open offer may be consistent with a nonconstant cost technology.

Since the  $z_t^i$  are normally distributed as  $N(0, \sigma_p^2)$ , the  $v_t^i$  are also normally distributed as  $N(\bar{p}, \sigma_p^2)$ .<sup>3</sup> We will assume that both the purchaser and suppliers are aware that the  $v_t^i$ s follow a normal distribution, but are ignorant of the true cost parameters  $\bar{p}$  and  $\sigma_p^2$ .<sup>4</sup> In fact, it is this ignorance that makes information acquisition a useful economic activity in this market.

Before continuing, we will without loss of generality (see, for example, Charles Plott, 1980) reorder the valuations  $v_t$  and the bids  $b_t$  such that  $v_t^i$  represents the  $i$ th lowest valuation, and  $b_t^i$  the  $i$ th lowest bid.

The fact that costs are stochastic, along with our assumption that projects awarded in adjacent periods are partial substitutes for one another, leads the purchaser to intertemporally allocate his demand in such a way that he buys a lot in periods when the low bid is unusually low, reflecting low costs, and buys less when the low bid is unusually high, reflecting high costs. In particular, the purchaser's choice of  $Q_t$  depends not just on the period  $t$  unit price,  $b_t^1$ , but also on the expected price in period  $t+1$ , which we denote  $e_{t+1}$ . Hence  $Q_t = Q(b_t^1, e_{t+1}; Q_{t-1})$ .<sup>5</sup> When  $b_t^1$  rises relative to the future price

$e_{t+1}$ ,  $Q_t$  falls, so that  $Q_1 < 0$ ; conversely, when the future price  $e_{t+1}$  rises relative to  $b_t^1$ ,  $Q_t$  rises; hence  $Q_2 > 0$ . Moreover, we will assume that  $Q$  is bounded for all pairs  $(b_t^1, e_{t+1})$ , and that there exists a bid  $b_t^0$  such that  $Q(b_t^0, e_{t+1}; Q_{t-1}) = 0$ .

In addition to choosing  $Q_t$  in period  $t$ , the purchaser must also decide whether or not to expend the fixed costs necessary for holding an auction in period  $t+1$ . This decision,  $P_{t+1}$ , will in part depend upon how much the purchaser hopes to buy next period, which in turn depends upon both  $e_{t+1}$  and  $Q_t$ ; that is,  $P_{t+1} = P(e_{t+1}, Q_t)$ . We assume  $P_1 < 0$ ,  $P_{11} < 0$  ( $P$  convex in  $e_{t+1}$ ),  $P_1(b_t^1, Q_t) = 0$ ,  $P_2 < 0$ , and that some  $x$  exists such that for all  $e \geq x$ ,  $P(e, Q_t) = 0$ .

Thus far in the description of our auction market we have pointed out that stochastic fluctuations in cost create an incentive for purchaser intertemporal substitution, and that this intertemporal substitution depends upon the comparison of the current market price,  $b_t^1$ , to the expected future price,  $e_{t+1}$ . We would now like to further point out that it is these very same stochastic cost fluctuations that cause  $e_{t+1}$  to be uncertain, and thus provide incentive for our purchaser to extract the information contained in each period's bids in order to improve his estimation of  $e_{t+1}$ . In fact, much of the remainder of this paper will be concerned with how the purchaser forms this estimate, and the consequences of the estimate for the economic behavior of the suppliers.

Since the underlying cost variables in our market,  $\bar{p}$  and  $\sigma_p^2$ , are stationary over time, each period's bids will improve the purchaser's ability to estimate  $e_t$  by providing him with "another round" of information. Thus, we choose a Bayesian update mechanism as the most natural method of modeling the purchaser's estimation procedure. We define  $g_t \sim N(\theta_t, \sigma_t^2)$  (with cumulative distribution function  $G_t$ ) to be the purchaser's best approximation to the cost density  $f$  following the submission of the period  $t$  bids, and  $\hat{N}_{t+1}$  to be the expected number of bidders in period  $t+1$ . We present the mathematical details of the Bayesian mechanism in the Appendix; however, we sum-

<sup>3</sup> Many of our results in fact apply to a much wider class of distributions.

<sup>4</sup> At this point it might be useful to distinguish between what we call production information and market information. For our purposes, the former is information on  $\bar{p}$  and  $\sigma_p^2$  represented by direct observation of the production process, while the latter is information on these parameters obtained from observations on market prices. In the case below we restrict our formal analysis to asymmetric market information—an asymmetry that in fact is produced by our cartel. While it is likely that there is a significant asymmetry in production information between the buyer and his suppliers, it is not required for our analysis of cartel behavior. Asymmetry in production information may, however, be useful, as we indicate in the introduction, in understanding the underlying structure of the problem.

<sup>5</sup> We are ignoring questions of risk and are assuming that the purchaser avoids the complexity of working the entire  $Q$  distribution. Note that  $Q_t$  also depends on  $Q_{t-1}$ , because when  $Q_{t-1}$  has been large,  $Q_t$  tends to be lower since the purchaser's cumulative demand over several periods is downward sloping.

marize the mechanism by the expression for  $e_{t+1}$ :

$$(2) \quad e_{t+1} = b\left(\hat{N}_{t+1}, \int_0^\infty x \hat{N}_{t+1} [1 - G_t(x)]^{\hat{N}_{t+1}-1} g_t(x) dx\right),$$

where  $b(\cdot)$  is a function relating the expected low valuation to the expected low bid. Note that  $b(\cdot)$  depends not only on the low valuation, but also directly on the expected number of bidders. This is the case because when bidders follow Nash noncooperative strategies in the auction, they will shade their bids upwards in an amount dependent upon the number of rival bids anticipated.<sup>6</sup>

It is clear from (2) that the expected number of bidders in period  $t+1$ ,  $\hat{N}_{t+1}$ , is a major determinant of the expected low bid,  $e_{t+1}$ . Much of the previous auction literature has exogenously specified the number of bidders in the market; we propose to improve upon this approach by determining  $\hat{N}_{t+1}$  endogenously in our model as a function of expected market demand,  $\hat{Q}_{t+1}$ . In fact, due to purchaser intertemporal substitution, expected long-run demand will often not equal expected short-run demand, and so we make a distinction between 1) the available pool of bidders in period  $t+1$ , which consists of suppliers who have made a long-run decision to enter the market, and 2) the number of suppliers out of this available pool who actually submit bids in period  $t+1$ .

The available pool of bidders consists of suppliers who have entered the market, and expect to remain for several periods. The equilibrium number in this pool will be a function of the long-run expected demand in the market as of period  $t$ ,  $Q(e_{t+1}, e_{t+1})$ , which represents demand at the price  $e_{t+1}$ , devoid of any short-run substitution effects. Specifically, the equation relating current

market conditions to the expected pool of bidders is assumed to be of the form:

$$(3) \quad \hat{R}_{t+1} = R_t + r(Q(e_{t+1}, e_{t+1}), R_t), \\ r_1 \geq 0, \quad -1 \leq r_2 \leq 0.$$

Equation (3) implicitly distinguishes between the behavior of the  $R_t$  suppliers already in the bidders pool and potential entrants. This is a valid distinction when there are assumed to be fixed costs of entry and exit.

We will presume that for each  $e_{t+1}$  there exists an equilibrium pair  $(Q^*, R^*)$  such that  $r(Q^*, R^*) = 0$ , and  $\hat{R}_{t+1} = R_t$ .<sup>7</sup> Furthermore, since entry or exit requires a fixed cost,  $r_1(Q^*, R^*) = 0$ . That is, small fluctuations in  $Q$  around its equilibrium value will not induce suppliers to change status. In specifying (3) we assume that while the industry is constant cost there is a minimum scale of operation. Hence given that "all firms are created equal," as  $Q(e_{t+1}, e_{t+1})$  rises, the number of bidders in the pool that can successfully survive also rises.

When there are fixed costs of determining the valuations  $v$ , the actual number of suppliers bidding in period  $t+1$ ,  $N_{t+1}$ , will generally not equal  $R_{t+1}$ , but instead will fluctuate because of purchaser intertemporal substitution. Specifically, if we summarize the period  $t+1$  demand curve as  $\hat{Q}_{t+1}$ , expected demand in period  $t+1$ , then the equation for the number of bidders expected in period  $t+1$  will be of the form:

$$(4) \quad \hat{N}_{t+1} = n(\hat{Q}_{t+1}, \hat{R}_{t+1}),$$

where  $n_1 > 0, n_2 > 0$ .

Equation (3) states that  $\hat{N}_{t+1}$  depends upon  $\hat{Q}_{t+1}$ ; we have already specified that  $\hat{Q}_{t+1}$  depends upon  $e_{t+1}$ , and that  $e_{t+1}$  in turn depends upon  $\hat{N}_{t+1}$ . Thus the three variables  $e_{t+1}$ ,  $\hat{Q}_{t+1}$ , and  $\hat{N}_{t+1}$  are interdependent and

<sup>6</sup>John Riley and William Samuelson (1981) have explicitly calculated the Nash function  $b(\cdot)$  and its embedded shading. It is of the form

$$b_i^j = v_i^j + \int_0^{v_i^j} F^{\hat{N}_i-1}(x) dx / F^{\hat{N}_i-1}(v_i^j).$$

<sup>7</sup>We say "equilibrium pair" because while  $Q$  affects the number of suppliers,  $R$  will also affect  $Q$  through its effect on  $N$  (see below).

must be simultaneously determined for our auction market to reach equilibrium in period  $t + 1$ .

Our auction market is now fully described, and the description thus far is an accurate portrayal of the market under competitive conditions. However, since the major point of our next section will be to show how a bidders' cartel can drive an informational wedge between itself and the purchaser, it is appropriate to mention the informational aspects of this market that are vulnerable to a cartel. First, the purchaser uses past bids to estimate the market's underlying cost parameters  $\bar{p}$  and  $\sigma_p^2$ , as well as the pool of bidders,  $R_t$ ; as a result manipulation of bids can bias his estimates. Second, equation (4) describes a short-run entry and exit equation that affects the future expected price,  $e_{t+1}$ . Since a cartel can control who bids, it can mislead the purchaser about the sensitivity of entry and exit to  $e_{t+1}$  by mimicking the form of (4) with a "bogus" equation of its own choosing.

## II. Theoretical Results

We have provided a set of behavioral equations that describe the functioning of our auction market under competitive conditions. Central to this description is the purchaser's use of the information contained in past auctions to improve his forecast of the future market price. To the extent that the purchaser is successful in applying the knowledge he gains from past auctions to the prediction of future price, he gains informational market power relative to the bidders in that he can more effectively manage his demand over time. As a result, the assumption that the auction market is competitive is suspect, because it implies that the auction's bidders collectively provide the purchaser with a "free lunch" consisting of their private valuations of project costs and market conditions. Economic theory suggests that such a free lunch will scarcely outlast breakfast. Rational bidders will find a means of either charging the purchaser a premium for their knowledge, or, preferably, consistently misinforming him and so preserving their informational advantage.

In this section we will demonstrate that, by forming a bidders' cartel, the auction participants can intentionally pass the purchaser misinformation about both project costs and the short-run entry and exit behavior of suppliers. Monopolizing market information and distorting the purchaser's view of the auction market will allow the cartel to achieve extraordinary profits, even by cartel standards, and will in addition mask the cartel's presence and so minimize the chances of the cartel being detected. In fact, as our argument proceeds, it will become clear that an uninformed purchaser's attempt to generate market information through the auction mechanism actually encourages the formation of a cartel.

We assume that all available bidders enter the cartel, and we concentrate on total cartel profits rather than the apportionment of these profits to specific cartel members. Our analysis of the bidders' cartel proceeds in several steps. We begin by showing that by manipulating the three variables  $R_t$ ,  $m_t$ , and  $s_t^2$ , the cartel can successfully alter the purchaser's perception of next period's low bid,  $e_{t+1}$ . Next we discuss the cartel's optimal choice of  $e_{t+1}$  and the low bid,  $b_t^1$ ; this discussion is divided into several parts, first an existence proof, then a demonstration of a standard period-by-period monopoly as a point of comparison with our more sophisticated cartel, and finally a detailed presentation of our cartel's short- and long-run strategies. Following this, we consider a more general cost scenario in which  $\bar{p}$  and  $\sigma_p^2$  stochastically range from period to period. As we shall see, this variability introduces additional noise into the auction market and works to the cartel's advantage. Finally, we suggest that by modifying his expectations from a fully rational structural model of the auction market to a simpler adaptive expectations mechanism, the purchaser can reduce his susceptibility to cartel exploitation.

### A. The Cartel's Ability to Manipulate $e_{t+1}$

In our multiperiod auction market, the purchaser extracts the information contained in past auctions' bids through a Bayesian mechanism and applies this information to

improved estimation of the expected future market low bid,  $e_{t+1}$ . Thus, the ability of a bidder's cartel to pass the purchaser misinformation about project costs and market conditions depends upon its ability to alter bids in such a way as to affect the purchaser's estimate of  $e_{t+1}$ . In our model the three variables through which the period  $t$  bids affect the purchaser's estimate of  $e_{t+1}$  are  $R_t$ ,  $m_t$ , and  $s_t^2$ . Clearly a cartel consisting of all available bidders can arbitrarily pick any or all of these variables in period  $t$ , since it controls who bids and what they bid.

Now suppose the cartel wishes to inflate the purchaser's estimate of  $e_{t+1}$ ; the following theorem shows that this can be accomplished by any one of the three strategies: (a) reducing the number of market suppliers  $R_t$ ,<sup>8</sup> (b) increasing  $m_t$ , or (c) decreasing  $s_t^2$ :

**THEOREM 1: Simple Cartel Information Strategies.** *The cartel can inflate  $e_{t+1}$  by (a) reducing  $R_t$ , (b) raising  $m_t$ , or (c) reducing  $s_t^2$ ; that is,  $\partial e_{t+1}/\partial R_t < 0$ ,  $\partial e_{t+1}/\partial m_t > 0$ , or  $\partial e_{t+1}/\partial s_t^2 < 0$ . The case in which the cartel wishes to deflate the purchaser's estimate of  $e_{t+1}$  is identical. (For proof, see Appendix II.)*

Theorem 1 demonstrates that the cartel can essentially manipulate  $e_{t+1}$  by manipulating  $R_t$ ,  $m_t$ , or  $s_t^2$ . Obviously the cartel can do the same by manipulating a combination of all three variables. In fact, since the cartel has at its disposal three "instrumental" variables with which to affect only one endogenous variable, it will, each period, have its choice from a wide range of alternative combinations of  $R_t$ ,  $m_t$ , and  $s_t^2$  in manipulating the purchaser's estimate of  $e_{t+1}$ .

Note that Theorem 1 forms the basis of the empirical tests of our model in the highway construction industry, which we present

in Section III. We have found strong evidence consistent with the view that highway bidders' cartels actually use these strategies to misinform government procurement agencies.

### B. The Cartel's Choice of $b_t^1$ and $e_{t+1}$

Having shown that the cartel can in effect "choose"  $e_{t+1}$  for the purchaser, we turn our attention to the cartel's maximization problem, its choice of an optimal  $b_t^1$  and  $e_{t+1}$ . To begin our analysis of this process we express total cartel profits from period  $t$  onwards as

$$(5) \quad \pi_t = Q_t(b_t^1 - v_t^1) + \sum_{i=1}^{\infty} \frac{P_{t+i}Q_{t+i}(b_{t+i}^1 - v_{t+i}^1)}{(1+\delta)^i},$$

where  $\delta$  is the cartel's discount factor. Since  $Q$  is bounded from above and below, and  $P$  effectively bounds  $e$  (i.e.,  $y$  exists such that for  $e > y$ ,  $P(e, Q) = 0$ ), the sum on the right of the right-hand side of (5) will be finite—we will denote it as  $J(b_t^1, e_{t+1})$ .

Formally, the cartel's maximization problem is

$$(6) \quad \text{Max}_{b_t^1, e_{t+1}} Q_t(b_t^1, e_{t+1}; Q_{t-1})(b_t^1 - v_t^1) + J(b_t^1, e_{t+1}).$$

Since  $Q$  is bounded, the implicit constant marginal cost technology of the suppliers implies that a well-defined interior solution exists to this problem:

**THEOREM 2: Existence Theorem.** *In period  $t$ , an optimal cartel choice of  $(b_t^1, e_{t+1})$  exists.*

(This is a standard nonlinear programming result; therefore, we omit the proof.)

In order to relate the behavior of our bidders' cartel to existing theories of collusion, we begin our analysis of (6) by discussing the limiting case in which purchaser decisions are made independently of  $e_{t+1}$ , which will be the case when there is no intertemporal substitution,  $\partial Q/\partial e = 0$ , and auctions

<sup>8</sup>In our formulation  $N_t$ , the actual number of bidders in period  $t$ , does not affect period  $t+1$ , whereas  $R_t$ , the number of suppliers, does, because of fixed entry and exit costs. However,  $R_t$  is in some sense unobservable, and must be extrapolated from the behavior of  $N_t$  over several periods. For a further discussion of this point see our empirical work in Section III.



are guaranteed to be held every period, so that  $P_t = 1$  for all  $t$ . In this situation, auctions held in different periods have no effect on one another, and so, in period  $t$ , the cartel faces a standard single-period monopoly market and will only be concerned with choosing the single variable  $b_t^1$ . Hence  $b_t^1$  will equal the standard monopoly price  $b_t^m$ , which will depend upon the unit marginal cost  $v_t^1$ , and the purchaser's demand function  $Q_t(b_t^1)$ . Formally, the problem is simply,

$$(7) \quad \pi_t = \text{Max}_{b_t^1} Q_t(b_t^1)(b_t^1 - v_t^1),$$

and the  $J$  function is independent of  $b_t^1$ .

When the purchaser's elasticity of substitution between periods is significant, and  $P_t$  is free to vary, a profit-maximizing cartel will not ignore the effect of  $e_{t+1}$  on profits  $\pi_t$ . In fact, by distorting the purchaser's perception of  $e_{t+1}$ , the cartel can earn significantly higher profits than the standard single-period monopoly case we have just outlined.<sup>9</sup> We will characterize this complete model, in which the cartel chooses both  $b_t^1$  and  $e_{t+1}$ , by considering the cartel's short- and long-run strategies separately.

### C. Short-Run Cartel Behavior

When the bidders' cartel first begins operating in our auction market, it will raise both  $b_t^1$  and  $e_{t+1}$  towards their monopoly levels. In the simple model we have presented, if the cartel were to set  $e = b^1$ , then it would face no risk either of detection or of the purchaser abandoning his structural model. In fact, however, the cartel will begin by inflating  $e_{t+1}$  above  $b_t^1$ , fooling the purchaser into believing that period  $t$ 's low bid is lower than period  $t+1$ 's low bid will be, and inducing the purchaser to substitute future demand forward into period  $t$ . Then, in period  $t+1$ , the cartel will inflate  $e_{t+2}$  above  $b_{t+1}^1$ , washing out the negative substitution from period  $t+1$ , again misleading the purchaser

into believing that the current market price is lower than the expected future market price. We will consider the "short run" to consist of the periods during which the cartel engages in this practice of setting  $e$ 's above  $b^1$ 's.

How long the short run lasts will depend on the cartel's perception of when the purchaser becomes suspicious about the discrepancy between  $e$  and  $b^1$ . We will concentrate on the first period.

In our model, during the cartel's first period it will set  $e$  near its upper bound, which is the value of  $e$  at which the purchaser is indifferent about holding next period's auction.<sup>10</sup> To see why increasing  $e_{t+1}$  above  $b^1$  increases profits, differentiate (6) with respect to  $e_{t+1}$ :

$$(8) \quad \frac{\partial \pi_t}{\partial e_{t+1}} = \frac{\partial Q_t}{\partial e_{t+1}}(b_t^1 - v_t^1) + \frac{\partial J}{\partial e_{t+1}}.$$

In the short run,  $e$  can be reset in period  $t+1$ ; as a result, all aspects of the  $J$  function will be independent of  $e_{t+1}$  except  $P_{t+1}$ , the decision about holding next period's auction, and  $\hat{Q}_{t+1}$ . Hence:

$$(9) \quad \left. \frac{\partial \pi_t}{\partial e_{t+1}} \right|_{SR} = \frac{\partial Q_t}{\partial e_{t+1}}(b_t^1 - v_t^1) + P_t \frac{\frac{\partial \hat{Q}_{t+1}}{\partial e_{t+1}}(b_{t+1}^1 - v_{t+1}^1)}{1 + \delta} + \frac{\partial P_t}{\partial e_{t+1}} \frac{\hat{Q}_{t+1}(b_{t+1}^1 - v_{t+1}^1)}{1 + \delta}.$$

Since  $P_t$  is convex,  $P$  will remain at 1 for small changes in  $e$ , hence  $\partial P_t / \partial e = 0$  for  $e$  near  $b^1$ . In addition, since  $e$  will be reset in  $t+1$ ,  $\hat{Q}_{t+1}$  depends on  $P_{t+1}$  only because of the constraint on total purchaser demand across several periods (i.e.,  $Q_t + Q_{t+1}$  must not exceed some bound on demand); as a

<sup>9</sup>The mechanism lying behind this claim is the cartel's ability to shift the purchaser's demand curve in or out by manipulating  $e_{t+1}$  in response to variations in cost.

<sup>10</sup>In a more complete model, the risk of detection will induce the cartel to choose  $e$  at some interior point between  $b^1$  and  $e$ 's upper bound.

result,  $\hat{Q}_{t+1}$ , which is discounted at  $1 + \delta$ , will never fall more than one-for-one with  $Q_t$ . Thus, (9) will generally be positive.

We now turn to the question: what is the cartel's preferred way of manipulating  $e$ ? In the short run, reducing the variance of bids is the most effective way of influencing  $e$ , because it not only directly lowers the purchaser's variance estimate  $\sigma_{t+1}$ , but also increases the weight the purchaser attaches to the current period's bids relative to his prior estimates. This implication that a cartel will sharply reduce the variance of bids in the short run provides the most direct link with our empirical tests below in North Carolina highway construction.

There are two ways in which our short-run cartel analysis could be significantly extended. First, one would like to predict when and where such a cartel will enter an auction market. Since the cartel must convince the purchaser of a one-time jump in costs, it will presumably pick a situation in which the purchaser has little past data that is relevant to the current auction.<sup>11</sup> Second, the question arises—what costs will the cartel incur should the purchaser abandon his structural model? We suggest some possibilities in Part F below.

#### D. Long-Run Cartel Behavior

As we have pointed out above, in the long run, our bidders' cartel will not be able to consistently inflate  $e$  above  $b^1$  and continually induce forward substitution. The purchaser will eventually realize that his method of estimation is incorrect rather than presuming, as he does in the short run, that he has merely misestimated parameters. In fact, in order to preserve the purchaser's faith in his structural model of the auction market, which is necessary if the cartel is to reap informa-

tional profits,<sup>12</sup> we will assume that the cartel must insure that the bids it submits fulfill the following consistency requirements.

First, on average, the purchaser's estimate of  $e_t$  must be unbiased, that is,

$$(10) \quad \lim_{T \rightarrow \infty} \frac{\sum_{t=0}^T b_t^1}{T} = \frac{\sum_{t=0}^T e_t}{T} = \bar{b}.$$

We note that, as in the short-run case,  $b_t^1$  will be chosen at a monopoly level, although not necessarily at the single-period monopoly price.

Second, since the market's cost and demand structures are known to be stationary, the purchaser's estimates of the mean, variance, and pool of bidders must eventually converge to constant values. Formally,

$$(11) \quad \lim_{\substack{T \rightarrow \infty \\ \varepsilon \rightarrow 0}} P\{|\theta_T - \theta_{T+1}| > \varepsilon\} = 0,$$

$$P\{|\sigma_T - \sigma_{T+1}| > \varepsilon\} = 0;$$

$$\lim_{T \rightarrow \infty} r \left( Q \left( \frac{\sum_{t=0}^T e_t}{T}, \frac{\sum_{t=0}^T e_t}{T} \right), R_t \right) = 0,$$

$$\hat{R}_{T+1} = R_T.$$

Third, whatever estimates the purchaser makes of the form of  $n(\cdot)$  must be consistently met by the market participants:

$$(12) \quad \lim_{T \rightarrow \infty} N_T = n_{\text{estimated}}(\hat{Q}_T, \hat{R}_T).$$

In fact, in the long run, the purchaser will expect the entry and exit equation (12) to fulfill a zero-profit condition. A bidder's expected profits are

$$\frac{\hat{Q}_{t+1}}{\hat{N}_{t+1}} (b_1 - v_1) - \text{fixed costs of submitting a bid.}$$

<sup>11</sup>For example, we would expect cartels to pick situations where there is 1) technological innovation which sharply reduces real costs; 2) a large shift in demand, such as a state decision to embark on a major road-building campaign; 3) a new product, such as the shift in road building to interstates; or 4) a substantial and sustained inflation.

<sup>12</sup>For a demonstration of this fact see Part F below.

If we assume that  $b_1 - v_1$  equals a constant,<sup>13</sup> then requiring that long-run expected profits be zero is equivalent to requiring that  $\hat{Q}/\hat{N}$  converge to a constant value:

$$\lim_{t \rightarrow \infty} (\hat{Q}_{t+1}/(\hat{N}_{t+1})) = \text{constant}.$$

We can now rewrite (12) in a simpler form:

$$(12') \quad \lim_{T \rightarrow \infty} N_T = \hat{N}_T = (\text{constant})(\hat{Q}_T).$$

Thus as  $\hat{Q}$  varies, due to purchasers intertemporal substitution,  $\hat{N}$  must vary proportionately. However, we note that since actual supplier profits are unobservable, the purchaser will not know what value the valid  $\hat{Q}/\hat{N}$  should equal: he will only know that in the long run it should have the same value every auction.

What sort of bids will the cartel submit in the long run? First, consider the choice of  $\bar{b}$ . Notice that by using its informational strategies the cartel can choose  $\sigma$  and  $\theta$  so that when this period's project is the long-run average  $e = \bar{e}$  and  $\bar{e} = \bar{b}$ . Thus, when the low bid is  $\bar{b}$ , the auction market is in a stochastic equilibrium, because next period's expected price is the same as this period's actual price. Thus the purchaser's demand function is  $Q(\bar{b}, \bar{b})$ . The cartel's choice of  $\bar{b}$  is now clear: it will pick  $\bar{b}$  to solve the monopoly problem of maximizing  $(\bar{b} - \bar{v})Q(\bar{b}, \bar{b})$ . We will denote the long-run demand induced by this  $\bar{b}$  as  $Q_A$ , and the expected number of bidders next period, when  $\hat{Q} = \hat{Q}_A$  also  $= Q_A$ , as  $\hat{N}_A$ .

Since the cartel ensures that  $\bar{e} = \bar{b}$ , the market will appear completely consistent with its competitive structure as modeled by the purchaser, and the cartel's presence will remain undetected. In contrast, a cartel that ignores the value of its informational strategies and inflates only  $b_i^1$  to its monopoly level, and not  $e_i$ , will in general fail to satisfy (10). To see this, note that in any period  $b_i^m$

must satisfy a pair of conditions:

$$(13) \quad e_i = b_i^m \quad \text{or} \quad \sigma_i Z_{N_i} + \theta_i = b_i^m;$$

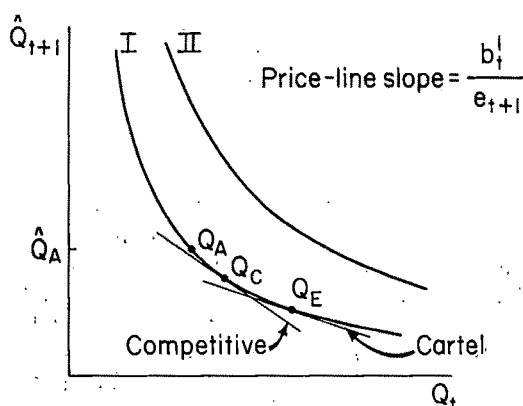
$$(14) \quad b_i^m = \text{Max}_{b_i^1} \pi_i,$$

where  $Z_{N_i}$  is derived from the standard normal power density of Appendix II. Since these conditions do not coincide, the  $b_i^m$  which satisfies (14) and maximizes profits will fail to satisfy (13), unless  $e_i$  is independently adjusted, and hence the average of the  $b_i^m$  over many periods will also fail to equal the long-run  $e$  average. In fact, a cartel which only adjusts  $b_i^1$  will find that  $e_i$ , which includes all bids, will consistently underestimate  $b_i^1$ ; the purchaser will then be tipped off to the presence of a cartel.

Our cartel's ability to hide itself is not its only long-run advantage. Specifically, we can investigate how our cartel bids when the low bid fluctuates around  $\bar{b}$ , due to fluctuations in cost around  $\bar{v}$ . Because unit costs are stochastic, our cartel will find that its unit profit margin is higher when  $b^1$  is below  $\bar{b}$  than when  $b^1$  is above  $\bar{b}$ . As a result the cartel has incentive to skew the purchaser's intertemporal allocation of demand so that he buys more in high-profit/low-cost periods, and less in low-profit/high-cost periods.

We will demonstrate that the cartel can in fact intertemporally reallocate the purchaser's demand in this manner. Note that the purchaser's choice of a pair  $Q_i$  and  $\hat{Q}_{i+1}$  depends on the relative price ratio  $b_i^1/e_{i+1}$ . Figure 1 depicts a set of  $(Q_i, \hat{Q}_{i+1})$  indifference curves and a typical  $b_i^1/e_{i+1}$  relative price line; given the current low bid  $b_i^1$ , the purchaser forms an estimate of  $e_{i+1}$ , and chooses  $Q_i$  and  $\hat{Q}_{i+1}$  at the point of tangency. What is unusual about this utility-maximization problem is that  $e_{i+1}$  itself depends upon  $\hat{Q}_{i+1}$ . Through (4)  $e_{i+1}$  is a function of the parameters  $\theta$ ,  $\sigma^2$ , and  $\hat{N}_{i+1}$  and while  $\theta$  and  $\sigma^2$  will settle down to constant values in the long run,  $\hat{N}_{i+1}$  will continue to fluctuate, according to (12), in response to fluctuations in  $\hat{Q}_{i+1}$ . As a result, the slope of the relative price line  $b_i^1/e_{i+1}$  will fluctuate as  $\hat{Q}_{i+1}$  moves. Specifically, when  $b_i^1$  is low,  $\hat{Q}_{i+1}$  will fall below its long-run average

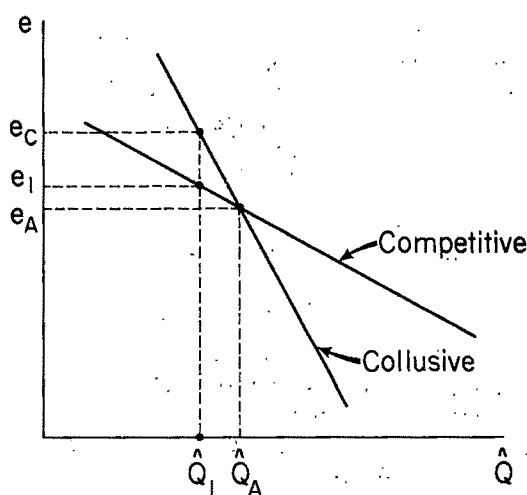
<sup>13</sup>Assuming that  $b_1 - v_1$  is a constant amounts to ignoring the dependence of shading on expected number of bidders. This effect is discussed in fn. 6.

FIGURE 1.  $(Q_t, Q_{t+1})$  INDIFFERENCE CURVES

value, denoted as  $\hat{Q}_A$  in Figure 1,  $e_{t+1}$  will then rise, and  $b_t^1/e_{t+1}$  will become flatter, decreasing  $\hat{Q}_{t+1}$  even more. As  $\hat{Q}_{t+1}$  falls,  $e_{t+1}$  then falls a bit more, further reducing  $\hat{Q}_{t+1}$ , and so on. However, this effect will quickly stabilize because, according to (4), the sensitivity of  $e$  to  $\hat{Q}$  rapidly lessens as  $\hat{Q}$  departs from  $\hat{Q}_A$ .<sup>14</sup> Conversely, when  $b_t^1$  is high and  $\hat{Q}_{t+1}$  is above  $\hat{Q}_A$ ,  $e_{t+1}$  will fall, further increasing  $\hat{Q}_{t+1}$ .

Suppose that the cartel was able to convince the purchaser that the sensitivity of  $e_{t+1}$  to  $\hat{Q}_{t+1}$  was larger than in the competitive auction. Then when a particularly low  $b_t^1$  caused  $\hat{Q}_{t+1}$  to fall, the purchaser's estimate of  $e_{t+1}$  would rise further than the competitive case, forcing  $\hat{Q}_{t+1}$  to fall further than the competitive case at the new tangency, and increasing  $Q_t$ . Figure 1 presents this case, the point  $Q_I$  corresponds to the cartel-induced tangency.  $Q_E$  corresponds to a larger  $Q_t$  than  $Q_C$  does, and hence to excess intertemporal substitution.

Figure 2 depicts the possible competitive and collusive  $e - \hat{Q}$  loci that will generate this result. The collusive locus is significantly steeper around the long-run stochastic equilibrium point  $(\bar{e}, \hat{Q}_A)$  than is the competitive locus.

FIGURE 2. THE RELATIONSHIP BETWEEN  $e$  AND  $\hat{Q}$ 

How might the cartel produce the steeper locus of Figure 2? We suggest two ways. First, the cartel can increase the long-run variance  $\sigma$ . The higher variance means that a given fluctuation in  $\hat{Q}$ , that causes through (12'), a fluctuation in  $\hat{N}$ , will generate a larger fluctuation in  $e$ . The cartel can also use  $\theta$  and  $\hat{N}_A$ , the long-run average number of bidders, to ensure that  $\bar{e}$  still equals  $\bar{b}$ .<sup>15</sup> Second, the cartel can manipulate the number of bidders. Recall that the purchaser does not know the slope in (12') relating  $\hat{N}$  to  $\hat{Q}$ . When this slope is steeper, the sensitivity of short-run entry and exit in the market to  $\hat{Q}$  increases, thus increasing the sensitivity of  $e$  to  $\hat{Q}$ . As an example of this, suppose that the monopoly solution consistent with the competitive  $\hat{Q} - \hat{N}$  slope has  $\sigma = 2$ ,  $\hat{N}_A = 4$ , and  $\theta_1$  chosen so that  $\bar{b} = 2$ . The cartel could then choose  $\theta_1^c$  so that  $\sigma^c = 2$ ,  $\hat{N}_A^c = 6$ , and  $\bar{b}$  again equals 2. Now consider a typical reduction in  $\hat{Q}$ . If  $\hat{N}_A$  falls by 1/2 to 2,  $\hat{N}_A^c$  will fall by 1/2 to 3. But  $(Z_6 - Z_3) > (Z_4 - Z_2)$  where  $Z_i$  is derived from the standard normal power density of Appendix II. Thus the cartel's change in  $e$ ,  $\sigma(Z_6 - Z_3)$ , will

<sup>14</sup> The alternative to this stabilization is a corner solution, which does not contradict our theory, although it is less interesting.

<sup>15</sup> From Appendix II,  $\bar{e} = \sigma Z_N + \theta$ . Neglect shading, when  $N = \hat{N}_A$ , choose  $\theta$  so that  $\sigma Z_{\hat{N}_A} + \theta = \bar{e}$ . The derivative with respect to  $\hat{N}_A$  is  $\sigma(\partial Z_N / \partial \hat{N}_A)$ , which increases linearly with  $\sigma$ .

exceed the competitive locus' induced change,  $\sigma(Z_4 - Z_2)$ .

Note that the cartel manipulates the purchasers estimate of the  $\hat{Q} - \hat{N}$  locus without violating the restrictions of his competitive model of the auction, which is summarized in (12'). As a result of this, the cartel will not reveal its presence. We thus conclude that with proper manipulation of either the variance or the number of bidders, the cartel can alter the long-run relationship between  $e$  and  $\hat{Q}$ , inducing excessive purchaser intertemporal substitution, and earning higher long-run profits than in the standard period-by-period monopoly case.

In summary, we have demonstrated an informational cartel's bidding strategy, and have shown that in the long run, an informational bidders' cartel will outperform more typical cartels on two counts. First, the information cartel will increase profits by increasing purchaser intertemporal substitution of demand; the cartel will accomplish this demand reallocation by manipulating how many of its members bid each period. Second, the cartel will earn these higher profits without ever revealing its presence, because it will not violate the purchaser's competitive structural model of short-run entry and exit in the auction. Specifically, by using its bids to control the purchaser's mean, variance, and pool of bidder estimates, the cartel will escape detection in the long run, while continuing to earn monopoly profits through inflated low bids.

#### E. Stochastically Varying Costs

Our model thus far has assumed the simple stationary cost structure given by equation (1). We conclude our discussion of cartel behavior by pointing out that when the uncertainty in production costs are more complex than this simple case, a wider informational gap can emerge between the bidders' cartel and the purchaser, allowing the cartel to increase profits even further.

As an example, consider the case when  $\bar{p}$  is known to be stochastically rising over time. When the purchaser is uncertain of the trend growth rate in  $\bar{p}$ , the cartel can inflate bids at a faster rate than costs, allowing its profits

to steadily increase. More commonly, however, the purchaser will relate the long-run price increase in the market to some index of economywide inflation, which will impose the constraint:

$$(15) \quad \lim_{t \rightarrow T} \bar{p}_t = a + bt,$$

where  $b$  = inflation rate;  $T$  = purchaser's perception of the long run.

In this situation, should the cartel have some knowledge of  $T$ , it can again increase profits by manipulating the "trajectory" of  $\bar{p}_t$  from the trend  $a + bt$  to a concave form in which prices rise rapidly at first, and then level off.

Finally, consider the situation where  $\sigma_p^2$  fluctuates in addition to  $\bar{p}$ . Here the market will resemble a "classic" rational expectations paradigm in which fluctuations in the low-bid market price must be differentiated into a permanent component, an increase in  $\bar{p}$ , and a transient component, a fluctuation in  $\sigma_p^2$ . We suspect that in this case it will be to the cartel's advantage to fluctuate  $\sigma_p^2$  more than costs dictate, thereby injecting noise into the market. The increase in noise will force the purchaser to attribute more of the fluctuations in market price to transient disturbances, and will therefore induce a higher rate of intertemporal substitution.

#### F. Purchaser Expectations

All of our results thus far have been obtained under the assumption that the purchaser uses a Bayesian mechanism to estimate  $e_t$ . We have been careful to point out that while the bidders' cartel will intentionally misinform the purchaser of the parameters of his structural model, it will whenever possible provide bids that are consistent with this structure. Should the purchaser abandon his Bayesian model of the auction market, the cartel may be discovered and/or lose its freedom to manipulate  $e_t$ .

The Bayesian structure we have presented uses the three variables  $R_t$ ,  $m_t$ , and  $s_t^2$  to estimate  $e_{t+1}$ . This would be the optimal approach if the purchaser faced competitive producers. Since the variables it uses are all "free" instruments under cartel control, they

provide the potential for extracting extraordinary profits from the purchaser. However, suppose that instead of the Bayesian approach, the purchaser uses a simple adaptive mechanism to calculate  $e_{t+1}$ . That is, suppose that, rather than calculating the mean, variance, and number of suppliers in the auction held during period  $t$ , and inserting these numbers into a structural model of next period's auction, suppose the purchaser merely forms a weighted average of past low bids in determining the expected future low bid. Specifically, suppose

$$(16) \quad e_{t+1} = \sum_{i=1}^{\infty} \alpha_i b_{t-i}^1 + \alpha_0 b_t^1.$$

In this case, short-run cartel profits will fall because the only means open to the cartel of manipulating  $e_t$  is the low bid  $b_t^1$ , and, if  $b_t^1$  is raised, the purchasers will substitute away from period  $t$ . Hence, the cartel suffers short-run losses as opposed to the short-run windfall as it attempts to raise  $e_t$ . More importantly, whereas in the rational expectations model the cartel can always exceed the standard period-by-period monopoly, in this case there is some uncertainty whether the cartel can even fully establish itself. The cartel must spend a long period of time convincing the purchaser to increase his expected price to the cartel's profit-maximizing level. Obviously, the longer the cartel has to spend inflating the purchaser expectations, the more likely the cartel is to fail. Thus one clear advantage of the rational expectations model for the cartel is the speed with which it can train the purchaser to treat the cartel price as "normal." We thus conclude that the adaptive expectations model constrains the cartel to approach the market's monopoly price gradually, since  $e_t$  can be raised above its competitive level only slowly. And while the cartel is struggling to attain its monopoly position, various factors may cause its disintegration because its short-run profits are in this case lower than the long-run case.<sup>16</sup>

<sup>16</sup> Clearly, extending our argument implies that when costs are more complicated than our simple stationary model, adaptive expectations will be even more of an improvement over the complete structural model.

### III. Some Empirical Observations

The most striking implications of our model of collusive behavior is that auction market cartels will actively pass misinformation to relatively uninformed purchasers. We have shown that, as long as purchasers use a rational expectations model to form price expectations, cartels will find it in their interest to manipulate more than simply the level of the low bid. We have identified three separate strategic variables through which a cartel can pass misinformation to a purchaser and hence manipulate the purchaser's estimate of the low bid. The three variables are 1) the mean of the bids submitted by cartel members; 2) the variance of the bids submitted by cartel members; and 3) the number of long-run market suppliers.

To empirically test the proposition that cartels do more than raise low bids requires a multiperiod auction market in which we can contrast competitive and collusive bidding patterns, and in which the prevalence of collusion was undetected for a significant period of time.<sup>17</sup> An example of a discovered cartel that meets these criteria is in highway construction. There has recently been an unprecedented level of federal antitrust activity in this sector. Between 1977 and 1982, the U.S. Department of Justice filed more than 200 indictments against highway contractors on charges of collusion. Prior to this period, however, antitrust activity in the industry had been rather quiescent and what activity there had been was quite geographically concentrated.<sup>18</sup>

We note that the informational strategies enumerated above will be operative primarily in the short run, but that the highway

<sup>17</sup> In a market where collusion is thought improbable, the assumption that purchasers use a rational expectations model of price formation is fully consistent with the assumption that purchasers choose an efficient mechanism for forming price expectations.

<sup>18</sup> Specifically, from 1955 to 1964 there was only one filing by the Antitrust Division for bid rigging in highway construction, and the defendants in this case were acquitted. In 1972 there was a single successful filing in Illinois and in 1974 there were seven such filings against firms in Illinois. It was not until very late in the 1970's and early in the 1980's that filings in the industry became numerous and widespread.

construction industry has had a sufficiently inflationary cost pattern that the five years we are studying may be considered to be a succession of short runs. In addition, since in highway construction the cartel does not control all projects, collusive firms are continually having to convince the highway department, by using the available informational strategies, that their bids are "reasonable" and that costs have, in fact, taken a once and for all jump.

The purchasers of highway construction are generally government agencies and hence detailed records of each auction are available. In addition, according to interviews with convicted cartel members, the highway bidders cartels that were discovered have not contained all producers, nor have the cartels controlled all bids submitted by members.<sup>19</sup> Hence, our sample contains both collusive and noncollusive bids, and we should be able to contrast cartel-controlled bid situations with competitive ones to determine if the cartel tried to manipulate the purchaser's perceptions of the costs of highway construction in accordance with the predictions of Theorem 1.<sup>20</sup>

Because of data availability we restrict our empirical analysis to the state of North Carolina.<sup>21</sup> In order to assure the plausibility of our assumption that the purchaser (in this case NC DOT) did not consider collusion a major problem, we further restricted the sample to projects put out to bid before 1980 and hence before most of the indictments were filed in North Carolina. Nevertheless, the sample generated by these restrictions is quite adequate since it contains about 1200

highway contracts and 25 colluding contractors.

The North Carolina Department of Transportation (NC DOT) classified each of the contracts in the sample as either collusive or competitive on the basis of conversations with apprehended colluders. Approximately 45 percent of the contracts were labeled collusive by NC DOT. In order to analyze the data we have constructed the indicator variable *COLLUDE* to represent this classification, where *COLLUDE* is 1 if the bids on the project reflected collusion and 0 otherwise.<sup>22</sup> It is important to note that because the classification is based exclusively on conversations with bid riggers, our measure of collusion is completely independent of the actual investigative process by which the officials in North Carolina actually discovered the cartel in the state.<sup>23</sup> This is a point that assumes some significance once we begin to analyze the relationship between this measure of collusion and the actual strategic variables developed below.

To assess the extent to which highway cartels in North Carolina adopted the strategies implied by our model, we first had to develop empirical measures of these strategies. The first two strategies (misinforming the purchaser as to the mean and variance of bids) relate to the bids submitted on a particular contract.<sup>24</sup> However, before the bids collected by NC DOT can be used to test our model, they must be modified.

<sup>22</sup> Despite NC DOT's claim that it has apprehended and questioned all colluders, *COLLUDE* may still reflect some erroneous classifications. Such errors tend to bias our analysis against finding significant differences in the patterns of collusive vs. competitive bid situations.

<sup>23</sup> According to Senior Deputy Attorney General Smith of North Carolina, in a personal communication, "essentially all labeling of NC highway contracts as collusive was based on conversations with bid riggers; essentially no contracts were labeled using investigative tools."

<sup>24</sup> We ignore the distinction between the bids and valuations in our empirical analysis. While we have assumed that we receive a random sample of bids that consequently reflect a random sample of valuations, the relationship between bids and valuations depends on the number of bidders anticipated. Since the anticipated number of bidders is unobserved, we cannot easily transform bids into valuations.

<sup>19</sup> Interviews with convicted cartel members in North Carolina yielded data both on cartel membership and on which contracts the cartel tried to rig. Interestingly enough, cartel members do not always win rigged contracts since there is some entry on specific projects.

<sup>20</sup> Since the cartel does not contain all potential producers, the cartels' bid patterns will also aim to distort the information of noncartel members.

<sup>21</sup> North Carolina provided us with computer readable data describing all highway contracts let in the state during the period 1975–81. The North Carolina data contains a project identifier, each bidder and his bid, the North Carolina Department of Transportation engineer's estimate of the project's worth, and assorted project characteristics.

The bids in this data set represent highway construction projects that vary widely in size, from \$10,000 to several million dollars. The predictions of the theory, of course, refer to bids on a similar size project. In order to handle this problem, we simply normalized the bids in the NC DOT data set by dividing each bid by the state engineer's estimate of the project's worth.<sup>25</sup> Of course, since the engineer's estimate is an estimate of how much the project should cost, dividing it into the bids, especially the low bid, does provide an indicator (at least in this short run) of how enhanced profits are on a specific project. Moreover, since the normalization in part reflects the profit level on the contract and profit levels vary systematically with the business cycle, this normalization is likely to be subject to cyclical variation.<sup>26</sup> This would appear to be a particularly bothersome phenomenon in our case, since the construction industry is notoriously cyclical. To correct for this second problem we have regressed the normalized low bid against a measure of unemployment in highway construction.<sup>27</sup> We have used this regression to adjust not

just the low bid, but all the bids; we denote these residuals as  $RESB_i^l$ .

The residual  $RESB_i^l$  is for the low bid and is likely (at least in the short run) to be a good indicator of cartel activity, that is, a good indicator of enhanced profits. The mean of the  $RESB_i^l$  is denoted  $\overline{RESB}$  and the variance is denoted  $CVBID$ .<sup>28</sup>

The third strategy requires cartel members to choose a bidding pattern that will skew the purchaser's estimate of the number of long-run market suppliers,  $R_t$ . While in most cases not all bidders for a job are cartel members, the cartel can control when its members bid, and how often they bid together. To understand how this might be used to manipulate the purchaser's sense of the magnitude of  $R_t$ , we present a simple recapture model that reflects how the purchaser might use bid patterns to estimate  $R_t$ .<sup>29</sup> Suppose the purchaser has no knowledge of  $R_t$ . He holds an auction and receives  $N_t$  bids. In the next period he holds an auction and receives  $N_{t+1}$  bids where  $K_{t+1}$  of the bids come from firms that also bid in period  $t$ . A naive estimate<sup>30</sup> of  $R_{t+2}$  would be  $N_t + N_{t+1} - K_{t+1}$ ; an estimate that suggests that in order to reduce the purchaser's estimate of  $R_{t+2}$  the cartel will want to choose  $N_t$  and  $N_{t+1}$  small, and  $K_{t+1}$  relatively large. For a variety of reasons, the cartel may not want to choose  $N$  too small;

<sup>25</sup> This method of normalization may bias the data against the effects we expect to find since the engineer's estimate may tend to be inflated on collusive jobs, lowering the value of the normalized bids. If there is any systematic difference between the types of jobs that are colluded upon and the rest of highway construction, then the engineer's estimate for the collusive type of jobs will eventually incorporate collusive profit levels. This is, of course, the objective the cartel is trying to achieve by manipulating the information the purchaser receives.

<sup>26</sup> There are two reasons why a cartel would respond to cyclical variations. First, it needs to make its behavior look like competitive bidding. Second, the behavior of (potential) noncartel bidders forces it to adapt its behavior to the state of the general construction market.

<sup>27</sup> The measure of unemployment is the number employed monthly in construction in a state divided by the ratio of annual average employed in that state's construction sector to one minus the annual employment rate. We have carried this regression out on a national data set of highway contracts in our research on the deterrent effects of antitrust enforcement. In this national regression we determined a single slope common to all states, but separate intercepts for each state. Specifically for North Carolina, this regression yields the regression  $RESB_i^l = b_1/e_t - (.775 + .138CYCLE)$ .

<sup>28</sup> These calculations include data on noncartel member bids for projects where the cartel attempted to rig the auction. While simple, this treatment may mask the difference between truly competitive and collusive bid situations where large numbers of noncartel members bid. This further weakens the  $\overline{RESB}$  and  $CVBID$  correspondence between  $m_t$  and  $s_t^2$  the mean and variance of the valuations. The name  $CVBID$  was chosen because the use of the ratio of the bids to the engineer's estimate is similar to calculating a coefficient of variation.

<sup>29</sup> For a discussion of recapture models, see Norman Johnson and Samuel Kotz (1970, ch. 6).

<sup>30</sup> This estimate is a lower bound on  $R_{t+2}$  if we assume no firms exit between  $t$  and  $t+2$ . A better estimator of  $R_{t+2}$ , assuming the bidding process is like sampling from a hypergeometric distribution, is  $((N_t + 1)(N_{t+1}) / (K_{t+1} - 1))$ . The variance of this estimator depends inversely on  $N$ .



thus we will focus on constructing a measure of  $K_i$ .<sup>31</sup>

For each contractor we have constructed an index  $H(i)$  reflecting how frequently a particular contractor bids with a certain select group of other contractors, as opposed to bidding with a wide variety of other contractors. This index for contractor  $i$ ,  $H(i)$ , is defined as the ratio of the number of different contractors  $i$  had bid with over the period to the number of bidders, other than  $i$ , who have also bid on projects that contractor  $i$  has bid on.<sup>32</sup>

To construct a comparable index for each contract, rather than each contractor, we take the average of the indices for the contractors bidding on a project, and denote this contract index as  $GROUP$ .<sup>33</sup> Contracts for which the  $GROUP$  variable is large correspond to small value of  $K_i$  in the estimator of  $R_i$  we presented above. Thus, if the cartel strategy entails misleading the purchaser about  $R_i$ ,  $GROUP$  should be significantly different for collusive and competitive contracts.

In the first column of Table 1 we present a set of bivariate regressions that relate  $COLLUDE$  (independent variable) to the empirical measures associated with the three cartel informational strategies outlined in Theorem 1. None of the intercepts are reported for the sake of brevity. The results are

TABLE 1—BIVARIATE REGRESSIONS BETWEEN INDICATORS OF CARTEL STRATEGY AND INDICATORS OF COLLUSION

Dependent Variables	Independent Variables <sup>a</sup>	
	<i>COLLUDE</i>	<i>INDRES</i>
$\overline{RESB}$	.030 (3.56)	—
$\overline{RESB}$	—	.124 (15.7)
<i>CVBID</i>	-.020 (-8.23)	—
<i>CVBID</i>	—	-.024 (-9.32)
<i>GROUP</i>	+.746 (-17.7)	—
<i>GROUP</i>	—	-3.81 (-8.00)
Number of Observations	1135	1135

<sup>a</sup>The coefficients divided by estimated standard errors are shown in parentheses.

remarkable support for our model of cartel behavior. Every coefficient is signed according to the predication of our theory and is statistically significant. The mean ( $\overline{RESB}$ ) of submitted bids are higher on bids labeled collusive,<sup>34</sup> the variance of the bids (*CVBID*) are less on collusive bids and the frequency with which bidders on a project have bid together (measured by the relative smallness of *GROUP*) is greater for collusive than noncollusive bids.<sup>35</sup>

<sup>31</sup>First, there are institutional reasons why  $N_i$  cannot be made small. Many states require a minimum number of bidders for projects of various sizes. Second,  $N_i$  plays a role in modifying the purchaser's estimate of future low bids since it determines the weight which the  $i$ th bids' mean and variance receive in updating prior information. For North Carolina, collusive bids have an average  $N$  of about 4 and noncollusive bid situations average about 4.5.

<sup>32</sup>As an example of the construction of  $H(i)$ , suppose two contracts are let, and contractors  $A$ ,  $B$ ,  $C$ , and  $D$  bid on the first contract, and contractors  $A$ ,  $B$ ,  $C$ , and  $E$  bid on the second; then  $H(A) = 4/6$ .

<sup>33</sup>If cartels form among those firms capable of highly specialized work, such as tunnels, we would expect those firms to be bidding together regularly even if they were bidding competitively. Organizing a cartel among a small number of well-known competitors should be relatively inexpensive and so may provide a second rationale for this indicator of collusion. We will discuss a partial test of this proposition later.

<sup>34</sup>It is interesting to note that the results relating  $\overline{RESB}$  to collusion do not simply reflect the effect of  $RESB^1$ , the low bid, on the mean of the bids ( $\overline{RESB}$ ). The results relating  $COLLUDE$  and  $\overline{RESB}$  are essentially the same when the low bid is omitted from the calculation of  $\overline{RESB}$ . Another indicator of the independent manipulation of  $\overline{RESB}$  by the cartel is the relationship between  $\overline{RESB}$  and  $RESB^1$  in this sample. The bivariate regression between  $\overline{RESB}$  and  $RESB^1$  is  $\overline{RESB} = .118 + .460 RESB^1$  (21.5), where the number in parentheses is the coefficient divided by its estimated standard error. Since the average number of bidders on a collusive contract in North Carolina is four, we would expect a coefficient on  $RESB^1$  of approximately .25 if it was only the low bid that was moving the mean.

<sup>35</sup>The two potential "flies in the ointment" here are: 1) the possibility that what we are observing is not the influence of collusion on strategic variables, ( $\overline{RESB}$ ,

In the second column of Table 1 we repeat the same bivariate regression with an alternative indicator of collusion. Here we classify the bids as collusive or noncollusive, not on the basis of whether they have been labeled as such by NC DOT, but rather by the magnitude of  $RESB^1$ . As we noted above,  $RESB^1$  is an indicator in the short run of the degree of profit enhancement embedded in the low bid. The variable  $INDRES$  represents the classification of projects by the size of  $RESB^1$ . The variable  $INDRES$  is 1 if the value of  $RESB^1$  is positive and zero otherwise.<sup>36</sup> While this indicator of collusion is just that, an indicator, and not an actual identification of collusive contracts, it is useful. One potential problem with the variable  $COLLUDE$  is that it undoubtedly involves some misclassifications, especially omissions

of contracts that were rigged but were not identified as such. The variable  $INDRES$  is unlikely to have the same classification bias. Moreover, the use of  $INDRES$  eliminates any possibility that the strategic variable  $GROUP$  is also the variable that generates the classification.<sup>37</sup> As is apparent from the results in Table 1, using  $INDRES$  as an indicator of collusion does not alter the results at all. All signs are as expected and all coefficients are significant.<sup>38</sup>

Another and perhaps somewhat more powerful approach to testing our model of cartel behavior is to consider how well our indicators of strategy jointly predict collusion. If our model is correct, then the probability that a contract is collusive should increase in the mean of the bids ( $\overline{RESB}$ ) and decrease in both the variables  $CVBID$  and  $GROUP$ . A direct test of the power of our model when all three strategies are considered simultaneously is presented in Table 2. We include  $RESB^1$  in this logit estimation basically as a control to see if the other strategic variables have independent predictive power. Again, the result of this empirical test conforms reasonably well to the predictions of our model.<sup>39</sup> When considered

$CVBID$ , and  $GROUP$ ), but rather the mechanism by which the collusive contracts in the sample were identified; or 2) that  $\overline{RESB}$ ,  $CVBID$ , and  $GROUP$  are simply proxies for the cost of collusion. If the collusive contracts were identified by the level of bids, or by the variability of bids, or by the degree to which contractors bid together, we might be observing the classification scheme instead of the impact of cartels on the levels of these variables. The fact that the collusion projects were actually identified by conversations with convicted bidders and not investigative techniques makes this much more a potential than actual problem in all but the case of  $GROUP$  and even here the possibility is quite remote (see fn. 37). On the second point that  $\overline{RESB}$ ,  $CVBID$ , and  $GROUP$  are reflections of the cost of collusion and not necessarily evidence of informational strategies being used by the cartel, we note that our empirical results are basically unaltered when we control for type of contract. If some types of contracts are more easily colluded upon because they represent a recurring and quite standard type of project, then perhaps measures such as  $CVBID$  and  $GROUP$  are simply indicators of the ease of collusion. The invariance, however, of our empirical results to type of project, at least as recorded by NC DOT, suggests that this is not the case.

<sup>36</sup>Because in the short run the purchaser's expectations haven't been fully "inflated" and  $RESB^1$  is a collusive bid,  $RESB^1$  will, in the short run, be positive for the cartel. Moreover, since in our sample not all contracts are collusive,  $e_i$  will reflect both collusive and noncollusive information and  $RESB^1$  will tend to be positive for collusive contracts and negative for competitive contracts. The latter is the case simply because the purchaser's estimate  $e_i$  is contaminated by the cartel's misinformation and hence a competitive bid will generally come in under the estimate.

<sup>37</sup>While  $GROUP$  was not used by NC DOT to identify bid-rigging, it might be acting as proxy for the convicted bid-riggers' memory in naming collusive contracts. The use of  $INDRES$  eliminates the possibility that the correlation between collusion and  $GROUP$  results from the classification process.

<sup>38</sup>It is interesting to note that the magnitude of the coefficients on  $CVBID$ , perhaps the most intriguing of the informational variables, is almost identical in both regressions.

<sup>39</sup>It is surprising that there is enough independent variation in the  $RESB^1_i$  and  $\overline{RESB}$  to support separate, statistically significant coefficients for both variables. This may reflect the fact that even rigged auctions are not entirely under the control of the cartel since they cannot keep noncartel members from bidding, and occasionally winning contracts. We have explored the issue of whether some of our indicators of collusive bidding are merely proxies for the type of project. We found similar results when we restricted the sample to particular types of jobs, such as those involving primarily resurfacing of existing highways. Similar results obtain if, rather than using  $COLLUDE$ , we classify contracts as collusive if they are won by a firm indicted for collusion.  $\overline{RESB}$  is insignificant in the multivariate model.

TABLE 2—MULTIVARIATE LOGIT RESULTS<sup>a</sup>

Independent Variables	Dependent Variable <i>COLLUDE</i>
<i>RESB</i> <sup>1</sup>	1.99 (2.27)
<i>RESB</i>	3.46 (3.18)
<i>CVBID</i>	-53.9 (-5.83)
<i>GROUP</i>	-22.0 (-13.5)

<sup>a</sup>The estimated asymptotic standard errors divided into the coefficients are shown in parentheses.

jointly, the three indicators of strategy *RESB*, *CVBID*, and *GROUP* are all related to the possibility of collusion in the manner suggested by our model and, more significantly, they show up as independent indicators of collusion even where we control for what is perhaps the best indicator of cartel activity in the short run, *RESB*<sup>1</sup>.

As noted above, in the short run, the cartel is in the process of manipulating the purchaser's expectations and its low bid will be above the purchaser's estimate of the project's cost. Moreover, since our sample contains both collusive and competitive contracts, the low bids on competitive contracts will tend to be below the purchaser's estimate of costs. This will be the case because while the purchaser's estimate is contaminated by data from collusive contracts, the bid itself in competitive contracts is simply a reflection of underlying costs. Hence, as we've noted previously, the magnitude of *RESB*<sup>1</sup> itself can be used in the short run as an indicator of the likelihood of collusion on a specific contract.

To the extent that the magnitude of *RESB*<sup>1</sup> is an indicator of collusion, our model suggests that it, like the direct indicator of collusion, *COLLUDE*, ought to be predictable from the value of *RESB*, *CVBID*, and *GROUP*. A test of our assertion concerning the predictability of *RESB*<sup>1</sup> is provided in Table 3. The results again are remarkably consistent with the implications of our theory. Moreover, since the dependent variable here is the magnitude of the low bid (*RESB*<sup>1</sup>)

TABLE 3—MULTIPLE REGRESSION RESULTS<sup>a</sup>

Independent Variables	Dependent Variable <i>RESB</i> <sup>1</sup>
<i>RESB</i>	.781 (31.2)
<i>CVBID</i>	-1.84 (-21.9)
<i>GROUP</i>	-.023 (-5.26)
Constant	-.014
Number of Observations	1135
<i>R</i> <sup>2</sup>	.53

<sup>a</sup>See Table 1.

and not whether the contract was identified as collusive, the results cannot simply reflect the fact that variables such as *RESB*, *CVBID*, or *GROUP* were used to identify collusion in the sample.

Overall, the empirical tests on the North Carolina data provide substantial support for our model. Cartels in practice appear to do more than simply raise the low bid on specific projects; they appear to be actively engaged in misinforming purchasers.

#### IV. Conclusion

In this paper we have investigated the impact of asymmetric information on the character and extent of collusion in multiperiod auction markets. Our analysis suggests that cartels will be especially pernicious in markets where information transmittal is potentially important. In such markets, cartels are likely to seriously impede if not entirely block the flow of accurate information through the market. Cartels in these cases appropriate the returns to superior information, and they do this not by disseminating the information, but rather by suppressing it. Cartels restrict not only the supply of traditional commodities, but also the flow of accurate information.

The passing of misinformation to the purchasers would appear, from our analysis, to be an extremely important aspect of cartel behavior, especially cartels operating in auc-

tion markets. In this regard, it is quite significant that we are actually able to observe this behavior in what is one of the most collusion-prone industries, road building. The empirical evidence that we have been able to assemble on bid-rigging in North Carolina road building indicates that cartels do in fact attempt to misinform purchasers. The evidence here indicates that cartels do more than simply raise the minimum bids on the specific projects.

The specific problem we chose for detailed analysis in this paper is a very common one. Essentially, we examined the problem of how a purchaser evaluates whether any specific offer represents a "good buy." While we study this problem in the context of an auction, the reader, we assume, will appreciate the generality of this concern. Judging whether a particular offer is a good buy or, in our case, whether a low bid is attractive, will be important to a purchaser whenever the prices facing him have an important stochastic component and whenever he has significant possibilities for intertemporal substitution.

#### APPENDIX I

Here we derive the Bayesian estimate of the expected future low bid,  $e_{t+1}$ . Let  $g_t$  denote the best subjective approximation to the true cost density  $f$  prior to period  $t$ 's auction. Following Fisher (see George Box and George Tiao, 1973), the equation relating  $g_t$  to the best approximation to  $f$  subsequent to period  $t$ 's auction,  $g_{t+1}$ , may be written

$$(A1) \quad g_{t+1}(x) = \lambda(\bar{p}, \sigma_p^2 | \text{data}) g_t(x).$$

Under competitive conditions the bids  $b_i^t$  form a Nash noncooperative equilibrium in the auction market such that no individual bidder has any incentive to deviate from his bid  $b_i^t$  when all other bidders follow their Nash strategies and bid  $b_j^t$ ,  $j \neq i$ . In general, the bids  $b_i^t$  are a function of both the number of bidders in the market,  $N_t$ , and the true valuations  $v_i^t$ . Given  $N_t$ ,  $b_i^t$  is known to be a strictly increasing, and hence invertible func-

tion of the valuation  $v_i^t$ .<sup>40</sup> furthermore, the function relating  $v_i^t$  to  $b_i^t$  can be explicitly calculated. Therefore, when the period  $t$  bids  $b_i^t$  are submitted, the period  $t$  valuations  $v_i^t$  can be exactly determined.

Armed with the bidders valuations  $v_i^t$  and the knowledge that the density  $f$  is normal, the update equation (A1) is particularly simple:<sup>41</sup>  $g_t$  should be chosen normal in each period; and, if  $g_t$  is distributed normal  $N(\theta_t, \sigma_t^2)$ , then  $g_{t+1}$  is distributed normal  $N(\theta_{t+1}, \sigma_{t+1}^2)$ , with

$$(A2) \quad \theta_{t+1} = (w_t \theta_t + N_t m_t) / (w_t + N_t)$$

$$\sigma_{t+1}^2 = \frac{w_t \sigma_t^2}{w_t + N_t} + \frac{w_t N_t (m_t - \theta_t)^2}{(w_t + N_t)^2} + \frac{N_t s_t^2}{w_t + N_t},$$

where  $w_t = w_{t-1} + N_{t-1}$ ,  $N_t$  = number of bidders in period  $t$ .

To estimate the expected low bid in period  $t+1$ , the revised density  $g_{t+1}$  is combined with the estimate of the expected number of bidders in period  $t+1$ ,  $\hat{N}_{t+1}$ , which is given by equation (4). The density of the expected low-bid valuation  $b_t^1$  is then

$$(A3) \quad \text{low}(x) = \hat{N}_{t+1} (1 - G_{t+1}(x))^{\hat{N}_{t+1}-1} g_{t+1}(x).$$

Finally, the expected low bid in period  $t+1$  is then simply the expectation of (A3) corrected for the distinction between the low valuation, and the Nash low bid:

$$(A4) \quad e_{t+1} = b(\hat{N}_{t+1}, \int_0^\infty x \hat{N}_{t+1} (1 - G_{t+1}(x))^{\hat{N}_{t+1}-1} g_{t+1}(x) dx).$$

<sup>40</sup> See, for example, Plott.

<sup>41</sup> For an exposition of this result see Box and Tiao (ch. 1).

## APPENDIX II

Here we prove Theorem 1:

- (a)  $\partial e_t / \partial R_t < 0$ ; (b)  $\partial e_t / \partial m_t > 0$ ;  
(c)  $\partial e_t / \partial s_t^2 < 0$ .

## PROOF:

In this proof we always assume that  $R_t$ ,  $m_t$ , and  $s_t^2$  may be chosen completely independently of one another, meaning that we consider the effects of any one of these three variables on the other two to be 0.

From (2),

$$e_t = b \{ \hat{N}_{t+1}, \bar{low}(\hat{N}_{t+1}, m_t, s_t^2) \}$$

where  $\bar{low}$

$$= \int_0^\infty x \hat{N}_{t+1} \{1 - G_{t+1}(x)\}^{\hat{N}_{t+1}-1} g_{t+1}(x) dx.$$

Thus

$$\begin{aligned} de_t = & \left( b_1 + b_2 \frac{\partial \bar{low}}{\partial \hat{N}_{t+1}} \right) \frac{\partial \hat{N}_{t+1}}{\partial \hat{R}_{t+1}} \frac{\partial \hat{R}_{t+1}}{\partial R_t} dR_t \\ & + b_2 \frac{\partial \bar{low}}{\partial \theta_{t+1}} d\theta_{t+1} + b_2 \frac{\partial \bar{low}}{\partial \sigma_{t+1}^2} d\sigma_{t+1}^2. \end{aligned}$$

From Appendix I and equations (3) and (4):

$$\frac{\partial \hat{R}_{t+1} \partial \hat{N}_{t+1}}{\partial R_t \partial \hat{R}_{t+1}} > 0;$$

$$d\theta_{t+1} = \frac{N_t}{w_t + N_t} dm_t > 0 \text{ for positive } dm_t$$

$$d\sigma_{t+1}^2 = \frac{N_t}{w_t + N_t}$$

$$\left( \frac{2w_t(m_t - \theta_t)dm_t}{w_t + N_t} + ds_t^2 \right) < 0$$

$$\text{for } ds_t^2 < \frac{-2w_t(m_t - \theta_t)dm_t}{w_t + N_t}.$$

From footnote 6 it follows directly that  $b_1 < 0$ ,  $b_2 > 0$ . Thus, we need only show that  $\partial \bar{low} / \partial \hat{N}_{t+1} < 0$ ,  $\partial \bar{low} / \partial \theta_{t+1} > 0$ , and  $\partial \bar{low} / \partial \sigma_{t+1}^2 < 0$ , and the results will follow.

Consider the standard normal density  $\phi(x) \sim N(0, 1)$ , and its associated cumulative distribution function  $\Phi(x)$ . Let  $Z_N = E\{\min(x_1, \dots, x_N)\}$  where the  $x_i \sim$  identically and independently distributed  $\phi(x)$ . Now consider  $g(x) \sim N(\theta_{t+1}, \sigma_{t+1}^2)$  and cumulative distribution function  $G_{t+1}(x)$ . Set

$$\bar{low} = E\{\min(y_1, \dots, y_N)\},$$

where  $y_i \sim$  identically and independently distributed  $g(x)$ . Then

$$\bar{low} = \int_{-\infty}^{\infty} x N g(x) [1 - G(x)]^{N-1} dx.$$

Now define  $\hat{x} = (x - \theta_{t+1}) / \sigma_{t+1}$ . Then

$$dx = \sigma_{t+1} d\hat{x}, \quad g(x) = \phi\left(\frac{x - \theta_{t+1}}{\sigma_{t+1}}\right) / \sigma_{t+1},$$

$$G(x) = \Phi\left\{(x - \theta_{t+1}) / \sigma_{t+1}\right\},$$

$$\text{and } xg(x) = [\hat{x}\sigma_{t+1} + \theta_{t+1}]\phi(\hat{x}) \frac{1}{\sigma_{t+1}}.$$

Substituting yields

$$\begin{aligned} \bar{low} &= N \int_{-\infty}^{\infty} \sigma_{t+1} \hat{x} \phi(\hat{x}) [1 - \Phi(x)]^{N-1} d\hat{x} \\ &\quad + \theta_{t+1} N \int_{-\infty}^{\infty} \phi_1(\hat{x}) [1 - \Phi(\hat{x})]^{N-1} d\hat{x} \\ &= \sigma_{t+1} Z_N + \theta_{t+1}. \end{aligned}$$

Using the well-known result (see Maurice Kendall and Alan Stuart, 1968, ch. 14) that  $Z_1 = Z_N < 0$  for all  $N > 1$ , and  $Z_N < Z_{N-1}$  for all  $N$ , it follows directly that

$$\frac{\partial \bar{low}}{\partial \hat{N}_{t+1}} < 0; \quad \frac{\partial \bar{low}}{\partial \theta_{t+1}} > 0; \quad \frac{\partial \bar{low}}{\partial \sigma_{t+1}^2} < 0.$$

(Note that  $\partial \bar{low} / \partial \sigma_{t+1}^2 < 0$  requires  $m_t > \theta_t$ , which follows from  $\partial \bar{low} / \partial \theta_{t+1} > 0$ .)

## REFERENCES

- Box, George E. P. and Tiao, George C., *Bayesian Inference in Statistical Analysis*, Reading: Addison-Wesley, 1973.
- Johnson, Norman L. and Kotz, Samuel, *Discrete Distributions*, New York: Wiley & Sons, 1970.
- Kendall, Maurice G. and Stuart, Alan, *The Advanced Theory of Statistics*, Vol. 1, New York: Hafner, 1968.
- Plott, Charles A., Jr., "Competitive Bidding for Contracts Under Alternative Auction Procedures," *Journal of Political Economy*, June 1980, 88, 433-55.
- Riley, John G. and Samuelson, William F., "Optimal Auctions," *American Economic Review*, June 1981, 71, 381-92.
- Stigler, George J., *The Organization of Industry*, Chicago: University of Chicago Press, 1968.

# Unemployment Duration and Incidence: 1968–82

By HAL SIDER\*

Cyclical changes in unemployment are a predominant feature of the labor market. Between 1973 and 1975, and again between 1979 and 1982, unemployment rates in the United States increased by roughly 70 percent. The labor market has further witnessed a large secular increase in unemployment over recent years. Between business cycle peak years 1969 and 1979, for example, unemployment rates increased by 65 percent. Such fluctuations and trends necessarily depend on the flows of workers into and out of unemployment. This paper examines the extent to which cyclical and long-term variations in aggregate unemployment reflect changes in the incidence of new spells and changes in average spell duration for those out of work.

A central issue in the analysis is the estimation of the length of completed unemployment spells from data on spells yet in progress. The construction and interpretation of such statistics has been at the heart of much recent debate on the dynamics of unemployment.<sup>1</sup> Previous duration estimates generally have been based on the assumption that unemployment reflects steady-state conditions; that is, that flows into and out of unemployment are constant over time. This

assumption is inappropriate for the analysis of cyclical fluctuations and long-term trends.

The paper first examines methodological issues in deriving estimates of the incidence and duration of completed unemployment spells. A modeling framework is developed and implemented based on aggregate probabilities that individuals continue in unemployment from one month to the next. The procedure does not rely on the restrictive steady-state assumption. The data that underlie construction of U.S. unemployment statistics for January 1967 through December 1982 are applied to the nonsteady-state model and are reapplied in the steady-state setting. Comparison of the estimates derived from these models indicates that calculations based on the steady-state framework systematically underestimate duration during recessions and overestimate duration at cyclical peaks. The paper then focuses on nonsteady-state measures of duration and incidence to explain fluctuations and trends in total unemployment. The results indicate that changes in duration play a very important role in explaining these phenomena.

## I. Measuring Unemployment Duration and Incidence

For many years, the Bureau of Labor Statistics has reported the average length of unemployment spells in progress. This measure, however, provides little direct evidence about the fully realized length of unemployment spells. Conceptually, the average duration of a completed spell can be determined by tracing an entering cohort through their unemployment experience.<sup>2</sup> The size of an entering cohort is denoted  $f(0)$  and the vector  $f(x)$ ,  $0 < x < n$ , represents the number of

\*U.S. Commission on Civil Rights, 1121 Vermont Avenue, NW, Washington, D.C. 20425. I thank John Cogan for many helpful discussions. I also thank Joe Antos, Arlene Holen, Michael Horrigan, Richard McDonald, George Neumann, John Raisian, Andy Sparks and the referees for their comments and assistance. Any errors, of course, are my responsibility. The research reported here was undertaken while I was at the U.S. Department of Labor. The views expressed are my own and do not necessarily represent the position or policies of the U.S. Department of Labor, or the U.S. Commission on Civil Rights.

<sup>1</sup>George Akerlof and Brian Main (1980, p. 885), for example, ask whether unemployment is better described in terms of stocks of people unemployed for long periods of time, or rather in terms of flows of persons whose spells of unemployment are quite short.

<sup>2</sup>See John Carlson and Michael Horrigan (1983) for a clear discussion of various duration measures and problems in interpreting these statistics.

individuals remaining in unemployment after each of  $x$  periods, where  $n$  is the maximum number of periods in unemployment. The average duration of a completed spell is simply the sum of these spells weighted by their completed length divided by the number of individuals that make up the cohort. In discrete terms, this can be written

$$(1) \quad S = \sum_{x=1}^n \frac{x(f(x-1) - f(x))}{f(0)} \\ = \sum_{x=0}^n \frac{f(x)}{f(0)}.$$

The process can be restated equivalently in terms of the probabilities of continuing in unemployment from one period to the next where  $p_x = f(x)/f(x-1)$ :

$$(2) \quad S = (1 - p_1) + 2p_1(1 - p_2) \\ + 3p_1p_2(1 - p_3) + \dots \\ = 1 + p_1 + p_1p_2 + p_1p_2p_3 + \dots,$$

or, more generally,

$$(3) \quad S = \sum_{x=1}^n g(x) \left[ \prod_{j=0}^{x-1} p_j \right] (1 - p_x),$$

where  $p_0$  is the probability of being in the initial cohort (and equals unity). The product of the  $p_j$ 's and  $(1 - p_x)$  is the share of the original cohort that exits unemployment after  $x$  periods. The function  $g(x)$  weights exiting individuals by the length of their completed spell. (In (2),  $g(x) = x$ ; although, as described below, this function can take a more general form.) Derivation of (3) does not require an assumption of steady-state conditions. Evaluation of (3) using current continuation rates yields an estimate of the expected spell duration for a synthetic cohort of individuals entering unemployment. It reflects the average completed spell duration that would be incurred if current continuation rates were maintained into the future. A variant of (3) based on this concept is estimated below.

The level of unemployment observed at any time depends on past and present values of  $f(0)$ , the size of entering cohorts, and past and present values for the vector of continuation rates. It is important to note, however, that if  $f(0)$  individuals enter unemployment each period and continuation probabilities are constant over time (i.e., the economy is in a steady state), then total unemployment at any point in time can be expressed:

$$(4) \quad U = f(0) + p_1f(0) + p_1p_2f(0) + \dots,$$

and

$$(5) \quad U = S \cdot f(0).$$

That is, under steady-state conditions, unemployment can be expressed as the product of incidence and average completed spell duration.

Implementation of nonsteady-state duration measures (such as (3)) requires estimates of continuation probabilities which can be derived only by observing the behavior of individuals or a cohort over time. However, if it can be assumed that unemployment reflects steady-state conditions, cross-section data on spells in progress can be applied to the problem of estimating completed durations.<sup>3</sup> In a steady state, the number of people leaving unemployment is equal to the number entering unemployment at any time and the duration distribution of spells in progress is constant. In such an equilibrium, differences in the observed number of in-progress spells of successive weeks duration reveal the probability that a member of the cohort remains unemployed an additional week. If steady-state conditions hold, mean completed spell duration can be calculated:

$$(6) \quad S' = \sum_{x=0}^n \frac{f(x, t)}{f(0, t)},$$

where  $f(x, t)$  denotes the number of individuals unemployed for  $x$  periods duration at

<sup>3</sup>This result was first discussed by Hyman Kaits (1970), who provides an excellent intuitive and algebraic derivation of the steady-state model.



time  $t$ . In a steady state, the number of individuals unemployed for  $x$  periods is constant so,  $f(x, t) = f(x)$  for all  $x, t$  and  $S' = S$ .

If steady-state conditions do not hold, then estimates based on this assumption will be biased. If unemployment is rising, then  $f(0, t) > f(0)$  and  $S' < S$ . In this situation, a cross-section distribution of in-progress spells by weeks duration is weighted too heavily by newer (and shorter) spells. The implicit continuation probabilities underestimate the true ones, resulting in underestimates of completed spell lengths.<sup>4</sup> When business conditions are improving, the cross-section distribution is weighted excessively by longer spells which results in overestimates of mean completed spell length. Thus, estimates based on inappropriate steady-state assumptions dampen actual fluctuations in duration over the business cycle.

Other authors have focused on the average completed spell length for the currently unemployed, as opposed to average duration for an entering cohort (George Akerlof and Brian Main, 1981). The completed spell length for this group is comprised of current plus remaining duration. Under steady-state conditions, persons are interviewed, on average, halfway through their unemployment spell, so simply doubling the average duration of spells in progress reveals the mean completed length of these spells. When steady-state conditions do not hold, however, the "doubling" estimator can be biased. The cyclical pattern of the bias is similar to that described above. In a recession, continuation

rates are rising and expected remaining duration on average exceeds the average current duration. A complementary bias is introduced when business conditions are improving.

Finally, steady-state methods can also bias estimates of the incidence of new spells. In a recession, the product of current incidence and mean completed duration exceeds the current level of unemployment because the newer, larger entering cohorts have not yet had time to filter through the entire duration schedule ( $U < S \cdot f(0, t)$ ). If average completed spell duration is known, steady-state estimates of incidence ( $I'$ ) calculated as the ratio of total unemployment to average completed duration (from (5)) will be biased downward in recessions ( $I' = (U/S) < f(0, t)$ ). The bias is reversed during periods of economic expansion. In this way, steady-state techniques may also dampen cyclical fluctuations in incidence. In practice, however, as discussed above, steady-state duration estimates can be biased which in turn can introduce an offsetting error in incidence estimates. In a recession, for example,  $S' < S$  so  $(U/S') > (U/S) < f(0, t)$ . The resulting cyclical nature of the bias in incidence estimates is ambiguous.

A large majority of attempts to analyze aggregate trends in unemployment duration and incidence have relied on steady-state techniques. These estimates have often been derived by fitting a smooth curve—often the *gamma* density—to average annual data on the number of spells in progress of various durations. The shape of this curve yields an implicit set of continuation rates for groups of individuals unemployed different lengths of time. The parameters of this function yield estimates of mean completed spell duration. This methodology, which essentially approximates equation (6), was developed by Hyman Kaitz (1970) and formalized by Stephen Salant (1977). Akerlof and Main (1980) applied these techniques to develop estimates of mean completed spell duration estimates for U.S. unemployment for 1948–78. Others, including Kim Clark and Lawrence Summers (1979) have implemented duration estimates based on continuation rates derived from longitudinal data (the gross-flows figures from the *Current*

<sup>4</sup>Consider the very simple example in which two individuals enter unemployment each period and the probability of remaining unemployed an additional weeks is .5. In equilibrium (weeks  $t, t+1$ ), a cross section of different cohorts yields continuation rates identical to those derived by tracing a cohort over time. If the size of an entering cohort increases, however, calculations based on a cross section of cohorts underestimate the true (constant) continuation rate and thus underestimate duration.

Duration	Number Unemployed		
	week $t$	$t+1$	$t+2$
1	2	2	3
2	1	1	1

Population Survey) and have thus avoided the steady-state assumption.<sup>5</sup> However, the data that form the basis of these estimates are generally not available in a time-series long enough to permit evaluation of secular and cyclical trends.<sup>6</sup>

## II. Data and Model Specification

Experiments with alternative methods for estimating the duration and incidence of unemployment spells are based on the monthly unemployment estimates of the Bureau of Labor Statistics.<sup>7</sup> Aggregate unemployment totals classified by reported weeks of spell duration (covering weeks 0–99) from January 1967 through December 1982 are utilized. These data reflect weighted counts from the *Current Population Survey (CPS)*. These schedules also serve as the basis of published unemployment duration schedules, which are reported in terms of intervals (less than 5 weeks, 5–10, 11–14, 15–26, 27+).<sup>8</sup>

The full schedule of in-progress spells is dominated by a pattern of spikes that reflect response bias among individuals in the sample. Local modes occur at durations corresponding roughly to monthly, quarterly,

half-yearly, and yearly points in the schedule. Few people, for example, report spells of five weeks duration relative to the number who report spells of four or six weeks. This pattern must be accounted for (and smoothed) in deriving either steady-state or nonsteady-state measures of completed spell duration. Consistent with this pattern of response bias, recent studies by James Poterba and Summers (1983) and by Norman Bowers and Francis Horvath (1985) reveal frequent errors in reported labor force status, unemployment duration and reasons for unemployment in the *CPS*. These insights are gleaned from the *CPS* reinterview surveys and matched longitudinal files on individuals who, due to the rotation group structure of the *CPS*, continue in the sample in consecutive months. The authors conclude that these problems may severely limit the reliability of econometric analyses focusing on matched longitudinal files. However, the offsetting nature of many reporting errors and the stability of rotation group bias tends to mitigate (though not eliminate) the impact of such errors on aggregate data, such as those utilized below.<sup>9</sup>

Problems and irregularities in the data have made the estimation of completed spell duration nearly as much an art as a science. Due to problems in fitting a smooth curve to the irregular weekly duration schedule, Kaiz and many subsequent analysts have utilized the

<sup>5</sup>Nicholas Kiefer, Shelly Lundberg, and George Neumann (1983) find that estimates of the duration distribution of completed unemployment spells derived from this type of data are sensitive to the assumed functional form. Although estimates of mean duration are fairly robust, alternative forms yield differing estimates of tail probabilities.

<sup>6</sup>An exception is J. K. Bowers and D. Harkess (1979) who estimated expected completed spell durations that rely on less-restrictive assumptions for biannual data from the British unemployment register for 1963–73. Norman Bowers (1980) has used the gross-flows data from the *CPS* to develop steady-state estimates of expected completed spell duration for 1969–79.

<sup>7</sup>Unemployment is defined to include individuals who are not working but have actively searched for a job in the past four weeks. The *CPS* questions on duration refer to the number of weeks since the onset of an individual's current (uninterrupted) spell. Unemployment spells can end with either employment or departure from the labor force.

<sup>8</sup>Implementation of these intervals in duration models incorporates the fact that, due to rounding procedures, reported intervals (< 5, 5–10, ...) reflect actual intervals (0–4.5, 4.5–10.5, ...). Alternative interval schedules outlined below also account for rounding rules.

<sup>9</sup>Both Poterba-Summers and Bowers-Horvath show that individuals in the *CPS* sample in consecutive months who report unemployment in both months tend to overstate increases in duration. This bias can affect the magnitude of duration estimates but does not necessarily affect cyclical duration patterns. For example, individuals classified as new entrants or reentrants to unemployment appear to overstate duration more than those classified as job losers and job leavers. However, the share of total unemployment made up by entrants and reentrants declines during recessions, a pattern that works counter to observed cyclical fluctuations in duration. Analysis of reinterview surveys by Poterba-Summers and Alfred Tella (1976) indicate that unemployment tends to be underreported as unemployed individuals are often classified as out of the labor force. Tella also finds more underreporting during recessions. The cyclical effect of such underreporting on duration estimates, however, is unclear as the location of such individuals in the duration schedule is not known.

published duration interval data as the basis of their estimates. The implementation of the nonsteady-state model (equation (3)) is also based on an interval representation of the underlying data, but the intervals are re-defined to correspond roughly to integer multiples of the monthly sampling window (0–4 weeks, 5–8, 9–12, 13–26, 27–39, 40–52, 53–75, 76–99).

With such an interval selection, a set of continuation rates can be derived by comparing adjacent interval populations in consecutive months. These continuation rates, in turn, are used to evaluate duration. For example,  $p_{1t}$  (the probability in month  $t$  of an individual continuing from his first to second month of unemployment) is calculated simply as the number of people in their second month of unemployment (at time  $t$ ) as a proportion of the number of individuals in their initial month of unemployment in month  $t-1$ :  $p_{1t} = h(1, t)/h(0, t-1)$ .<sup>10</sup> The aggregate probability of continuing in unemployment from the second to third month, and from the third to fourth month, etc., is calculated in a similar fashion. In this manner, a full set of continuation rates that exhaust the duration schedule can be derived for each month in the sample.<sup>11</sup>

Implementation of equation (3) further requires specification of  $g(x)$ , the function that weights individuals leaving unemployment by the appropriate completed duration. If spells

end, on average, halfway through the month then  $g(1)$  would equal .5;  $g(2)$  would equal 1.5; etc. Actually, the average departure date is somewhat before midmonth because as the month progresses, fewer members of the original group remain and the absolute size of the outflow declines. Specification of  $g(x)$  relies on the assumption that the probability of exit for individuals remaining unemployed is constant over the course of a month.<sup>12</sup>

Intervals selected as integer multiples of the sampling window have the property that the transition weeks in the schedule (weeks 4, 8, 12, etc.) generally correspond to spikes in the data. The published intervals fully incorporate transition weeks with adjoining intervals of shorter duration. This is not completely appropriate because spikes at transition weeks likely include individuals who underreport as well as some who overreport true duration. Accounting for this problem requires an algorithm to allocate transition points between adjacent intervals. The simple rule used below is to allocate these weeks equally across adjacent intervals.

Estimates of completed spell duration are evaluated using current continuation rates and weights,  $(p_{x,t}, g(x, t))$ . The resulting measure reflects the expected completed spell duration of a synthetic cohort that enters unemployment and faces current economic conditions throughout their unemployment spell. In this sense, the expected duration statistic permits estimation of long-run responses to steady-state changes in the macroeconomic climate.<sup>13</sup>

The assumptions required for construction of this statistic are less restrictive than those

<sup>10</sup> This contrasts with the implicit steady-state estimate:  $p_{1t} = h(1, t)/h(0, t)$ .

<sup>11</sup> More specifically, continuation probabilities are calculated through the following transitions:

< 5 weeks in month  $t-1$  to 05–08 weeks in  $t$   
 05–08 weeks in month  $t-1$  to 09–12 weeks in  $t$   
 09–12 weeks in month  $t-1$  to 13–16 weeks in  $t$   
 13–26 weeks in month  $t-3$  to 27–39 weeks in  $t$   
 27–39 weeks in month  $t-3$  to 40–52 weeks in  $t$   
 40–52 weeks in month  $t-3$  to 53–65 weeks in  $t$   
 53–75 weeks in month  $t-6$  to 76–99 weeks in  $t$   
 76–99 weeks in month  $t-6$  to 100+ weeks in  $t$

The 100+ category is defined as one-half of the total reporting 99 weeks, the truncation point in the data. These intervals more fully characterize the long-durations tail of the distribution of in-progress spells (relative to the BLS schedule) and slightly reparameterize shorter durations. Longer intervals at longer durations are necessary to smooth spikes in the data at 26, 52, 75, and 99 weeks.

<sup>12</sup> A similar point has been noted by James Luckett (1979). Assuming constant exit rates from each cohort within a month; one-half of eventual exits over the month have left when  $g(x) = \ln((1 + p_x)/2)/\ln(p_x)$ .

<sup>13</sup> Because calculation of long-duration continuation rates requires lags of more than one month, estimates of expected spell duration will tend to lag slightly current conditions. The bias, however, is likely to be relatively small due to the small share of spells that reach long durations and the relative stability of continuation rates at long durations. Expected duration measures have also been implemented by Clark and Summers (1979), Bowers and Harkess, and Bowers.

for the steady-state model, but are important nevertheless. More specifically, it is necessary to assume that steady-state conditions hold over the sampling interval. That is, it is necessary to assume that entrants to unemployment arrive at a constant rate and face constant exit probabilities throughout the interval period. This assumption, however, allows estimation of weekly inflows into unemployment. Specifically, the number of individuals observed with less than five weeks unemployment at time  $t$ ,  $h(0, t)$ , can be considered as the result of a renewal process that reflects a weekly continuation rate  $p_1^*$  (derived from the estimated monthly continuation rate for individuals in their first month of unemployment,  $p_1$ ) and the number of individuals,  $N$ , that entered unemployment each week over the course of the month. Assuming four weeks per month:

$$(7) \quad h(0, t) = N(1 + p_1^* + p_1^{*2} + p_1^{*3}).$$

This expression can be solved for  $N$ , yielding an estimate of the average weekly incidence of unemployment for each month of the sample.<sup>14</sup>

### III. Comparison of Alternative Estimates of Duration and Incidence

This section compares steady-state and nonsteady-state estimates of the completed duration and incidence of unemployment spells derived from aggregate unemployment data. First, steady-state duration estimates are derived using data defined in a manner that conforms with the nonsteady-state model; next, biases in steady-state measures are examined by comparing these estimates with nonsteady-state results.

The first step in the evaluation process is to examine the sensitivity of published steady-state results to changes in the parameterization of the duration intervals. The basic framework of Salant is reproduced utilizing a

nonlinear least squares algorithm to estimate the parameters of the *gamma* density from average annual data on in-progress spells. First, the results of Akerlof and Main, that were based on standard published duration intervals, are reestimated. The reproduced results (not reported) are very similar to those reported by the authors except for 1973, for which the recalculated spell duration is 71 percent of the original and for 1976, for which the model fails to converge.<sup>15</sup> For other years, recalculated mean durations average 98.5 percent of those reported by Akerlof and Main. The steady-state results are also robust with respect to respecification of the duration intervals. Reestimation on this basis results in mean completed spell durations that are, on average, 1.5 percent shorter than those derived using the published schedule.<sup>16</sup>

Allocating transition weeks between adjacent intervals, however, has a substantial impact on duration estimates. Because these transition weeks are no longer grouped with weeks of shorter duration, the revised distribution is less compressed toward shorter durations. This yields higher continuation rates and longer average durations. Allocating transition points equally between adjacent periods increases mean duration by an average of 34.1 percent. This change affects the levels but not cyclical patterns in the steady-state estimates. The simple correlation between annual measures of spell duration based on alternative treatments of transition weeks exceeds .99.

Before presenting the nonsteady-state estimates of expected completed spell durations, the set of continuation probabilities that underlie these estimates is examined (Table 1). The estimates reveal the oft-noted pattern that the aggregate probability of remaining

<sup>15</sup>The results of the analysis of the sensitivity of steady-state duration estimates are available on request from the author.

<sup>16</sup>This difference is attributable to the reparameterization of the shorter intervals, not the more detailed specification of the long-durations tail. Results based on a more detailed specification of the tail and standard shorter intervals are identical (to the tenth of a month) to the base results.

<sup>14</sup>A similar point is noted by George Perry (1972). Estimation of  $N$  is based on the average of 4.3 weeks per month.

TABLE 1—AGGREGATE MONTHLY PROBABILITIES  
OF CONTINUING IN UNEMPLOYMENT  
(Nonsteady-State Estimates)

Month of Unemployment	1969	1975	1979	1982
1	.41	.57	.51	.59
2	.51	.64	.55	.65
3	.59	.73	.63	.74
4-6	.61	.75	.64	.74
7-9	.72	.83	.78	.85
10-12	.94	.97	.94	.97
13-18	.92	.96	.89	.94
19-24	.90	.97	.92	.97

Note: Calculations described in text.

unemployed increases with duration. Continuation rates fall a bit at longer durations, a pattern that may reflect an increased likelihood of withdrawal from the labor force for

the very long-term unemployed. Continuation rates also follow a distinct cyclical pattern, rising during recessions (in 1975 and 1982, for example). This cyclical pattern is observed for each duration-specific group.

Annual means of completed spell durations estimated by steady-state and nonsteady-state methods are presented in Table 2. The reported estimates are based on identical interval specifications and treatment of transition points in the underlying data. The results show that nonsteady-state measures are, in general, slightly higher than steady-state estimates and that the magnitude of the difference is related to business cycle conditions. More specifically, the largest wedge between steady-state and nonsteady-state measures is observed in years of deteriorating business conditions (1970, 1975, 1982).

TABLE 2—ALTERNATIVE DURATION AND INCIDENCE MEASURES

	Expected Completed Spell Duration (weeks)		Weekly Inflow (thousands)	
	Steady State <sup>a</sup>	Nonsteady State <sup>b</sup>	Steady State <sup>c</sup>	Nonsteady State <sup>d</sup>
1968	6.1	6.1	462	439
1969	6.1	6.2	464	445
1970	7.0	8.2	584	561
1971	8.8	9.8	570	555
1972	8.3	8.7	588	561
1973	7.2	7.3	606	566
1974	7.4	8.4	697	660
1975	11.3	14.6	702	691
1976	10.5	11.3	705	667
1977	9.5	9.6	736	696
1978	8.3	8.3	747	701
1979	8.0	8.3	767	719
1980	9.5	10.5	804	771
1981	9.6	10.8	862	808
1982	11.4	14.3	937	902
Coefficients of Variation: <sup>e</sup>				
Unadjusted:	.191	.256	.192	.193
Detrended:	.135	.202	.041	.049
Cyclical Elasticities: <sup>f</sup>	-1.9 (.53)	-3.2 (.58)	-.40 (.26)	-.64 (.26)

<sup>a</sup>Derived from nonlinear least squares estimates of *gamma* density based on intervals (0-4, 5-8, 13-26, 27-39, 53-75, 76-99, 100+).

<sup>b</sup>Estimates of equation (3) based on same intervals.

<sup>c</sup>Derived from equation (5) using steady-state duration estimates.

<sup>d</sup>Derived from equation (7).

<sup>e</sup>Standard deviation divided by mean of series.

<sup>f</sup>Derived from regression of log of annual unemployment measures on *IPI*\* (deviations from trend in log of the Industrial Production Index). Standard errors are shown in parentheses.

This is consistent with the discussion in Section I.

As a result, the steady-state measure ( $S'$ ) exhibits considerably less variability over time than the nonsteady-state estimate ( $S$ ). For example, the coefficient of variation for the annual nonsteady-state model is .26 compared to .19 for the steady-state result. The magnitude of the bias introduced by steady-state techniques can be summarized in regressions of the log of annual duration measures on a trend term and business cycle indicator  $IPI^*$ , that is constructed as deviations from trend growth in the log of the Industrial Production Index. The results (also reported in Table 2) indicate that a 1 percent decline in  $IPI^*$  increases the nonsteady-state measure by 3.2 percent. The corresponding increase in  $S'$  is 1.9 percent. The difference between the estimates is significantly related to business cycle conditions.<sup>17</sup>

Despite the cyclical nature of the bias, in years of improving business conditions, the steady-state estimates do not exceed the nonsteady-state measure. One reason for this is related to the secular growth in unemployment that has resulted from increased growth in the labor force and apparent trend increases in the "full-employment" unemployment rate.<sup>18</sup> Steady-state methods are biased downward by trend growth in unemployment because the cross-section distribution is weighted too heavily by the newer and thus shorter spells. This growth accounts for a sizeable share of the "cyclically adjusted"

difference between steady-state and nonsteady-state measures.<sup>19</sup>

Table 2 also presents steady-state and nonsteady-state estimates of the incidence of new unemployment spells. The steady-state incidence measure is defined using the equality in (5) and the steady-state duration measure; the nonsteady-state incidence measure is estimated from (7). The two series are roughly similar in magnitude and variability. The steady-state measure is larger by an average of 5 percent and the nonsteady-state measure is somewhat more variable, although both measures are less variable than the duration statistics. Regressions of the log of annual incidence estimates on  $IPI^*$  and a linear trend indicate that a 1 percent decrease in  $IPI^*$  increases the steady-state measure by .4 percent and the nonsteady-state measure by .6 percent. The difference between the incidence measures is again significantly related to  $IPI^*$  indicating that steady-state methods dampen estimates of cyclical fluctuations in incidence.

#### IV. Trends in Unemployment Duration and Incidence

Analysis of secular and cyclical fluctuations in unemployment requires analysis of variation in its components. This section examines in a bit more detail secular and cyclical changes in unemployment based on the nonsteady-state measures of incidence and duration derived above. Trends in incidence and duration are described first; the long-run impact of business conditions on unemployment is then analyzed and the share of such fluctuations attributable to incidence and duration is examined.

Various unemployment indicators for 1968–82 are summarized in Table 3. Unem-

<sup>17</sup>A similar cyclical bias is observed in estimates of completed spell duration for the currently unemployed. The steady-state measure is calculated as twice average current duration of in-progress spells. A nonsteady-state counterpart can be calculated by adding the average current duration of in-progress spells to an estimate of expected remaining duration. The latter measure is constructed by evaluating an expected duration equation (similar in concept to equation (3)) for each cohort of the currently unemployed and then taking a weighted average of this measure over all cohorts. A 1 percent decline in  $IPI^*$  increases the "doubling" statistic by 1.4 percent and the nonsteady-state measure by 1.9 percent. The difference in these measures is significantly related to  $IPI^*$ .

<sup>18</sup>See Joseph Antos, Wesley Mellow, and Jack Triplet (1979) for a discussion and review.

<sup>19</sup>Business cycle "peak-to-peak" average monthly unemployment increased by roughly 8 percent annually between 1969 and 1979. Experiments with mean continuation rates were performed to simulate the size of various cohorts under this level of unemployment growth. Steady-state estimates based on the simulated data underestimate the true duration—which is constant by assumption—by about 3 percent, although this figure is somewhat sensitive to the assumed timing of the increase over the course of a year.

TABLE 3—UNEMPLOYMENT INDICATORS: 1968–82

	Unemployment Rate (1)	Average Unemployment (2)	Expected Completed Spell Duration (3)	Expected Completed Spell Duration for Currently Unemployed (4)	Weekly Inflow (5)	Incidence Rate (6)	<i>IPI</i> * (7)
1968	3.6	2822	6.1	18.2	433	.55	.005
1969	3.5	2835	6.2	17.4	440	.54	.022
1970	4.9	4092	8.2	20.1	555	.67	-.036
1971	5.9	4996	9.8	25.1	549	.65	-.046
1972	5.6	4843	8.7	25.0	554	.64	.015
1973	4.8	4306	7.3	20.6	559	.63	.069
1974	5.5	5078	8.4	21.3	653	.71	.039
1975	8.4	7827	14.6	32.7	684	.73	-.081
1976	7.6	7289	11.3	32.4	660	.69	-.007
1977	6.9	6857	9.6	28.3	689	.70	.024
1978	5.9	6049	8.3	24.1	693	.68	.053
1979	5.7	5965	8.3	23.1	712	.68	.069
1980	7.0	7452	10.5	26.4	763	.71	.006
1981	7.4	8080	10.7	29.2	801	.74	.005
1982	9.7	10681	14.3	35.9	903	.82	-.107

Source: Bureau of Labor Statistics for cols. (1) and (2).

Col. (2) Shown in thousands.

Col. (3) Nonsteady-state measures of expected duration for an entering cohort (shown in weeks).

Col. (4) Nonsteady-state measure for currently unemployed (shown in weeks).

Col. (5) Nonsteady-state measure of incidence (shown in thousands).

Col. (6) Incidence measure divided by average monthly labor force (shown in percent).

Col. (7) Deviations from trend growth in the log of the Industrial Production Index.

ployment and unemployment rates increased substantially over the period. This long-term increase is attributable as much to lengthened spell durations as to the increased incidence of unemployment spells. Between peak business cycle years 1969 and 1979, for example, average unemployment increased about 110 percent while the incidence of unemployment spells increased by only 64 percent.<sup>20</sup> The average unemployment rate increased 63 percent over these years while the incidence rate (calculated as the ratio of weekly incidence to average monthly labor force) grew by just 26 percent. The expected duration of unemployment spells increased by 34 percent over this period.

Cyclical changes in unemployment are composed of fluctuations in incidence and

average spell duration. Changes in duration, in turn, depend on variability in the underlying sets of continuation rates. A series of experiments is performed (similar to those reported in Table 2 for analyzing the cyclical bias in steady-state methods) in which the log of the continuation rates estimated for each month in the sample are regressed on a distributed lag of *IPI*\*, a time trend, and month-specific dummy variables.<sup>21</sup> The results indicate that continuation rates are more cyclically sensitive at shorter durations (see

<sup>20</sup>As a basis for comparison, average monthly initial claims for unemployment compensation increased by 94 percent between 1969 and 1979.

<sup>21</sup>The incidence model incorporates a 3-month distributed lag of *IPI*\*; the duration model a 6-month lag; and the unemployment and unemployment rate models are based on a 12-month lag. Lag lengths were selected by testing restrictions on models incorporating longer lags. Similar models were estimated by Clark and Summers (1981) to explain employment variability. Estimation included a correction for first-order serial correlation. Similar models were estimated utilizing BEA's index of coincident indicators to measure business cycle conditions and yielded very similar results.

TABLE 4—VARIABILITY IN UNEMPLOYMENT MEASURES, JANUARY 1968–DECEMBER 1982

	Mean	Coefficients of Variation		Regression Coefficients <sup>a</sup>	
		Unadjusted	Detrended, Deseasonalized	$\frac{\partial \ln Y}{\partial IPI^*}$	$\frac{\partial \ln Y}{\partial t}$
Expected Duration: <sup>b</sup>					
Entering Cohort	9.4	.36	.25	-3.19	.0028
Currently				(.74)	(.0006)
Unemployed	24.9	.24	.17	-1.98	.0029
				(.16)	(.0002)
Weekly Inflow <sup>c</sup>	632	.24	.07	-.60	.0035
				(.19)	(.0002)
Incidence Rate <sup>d</sup>	.67	.16	.07	-.70	.0015
				(.18)	(.0002)
Unemployment <sup>c</sup>	5774	.38	.17	-2.78	.0058
				(.63)	(.0018)
Unemployment rate <sup>d</sup>	6.0	.29	.18	-3.01	.0036
				(.55)	(.0009)
Continuation Rates:					
Month 1	.51	.16	.09	-.82	.0017
				(.12)	(.0001)
Month 2	.57	.13	.09	-.88	.0009
				(.11)	(.0001)
Month 3	.66	.13	.09	-.84	.0008
				(.13)	(.0001)
Months 4–6	.68	.09	.06	-.81	.0007
				(.08)	(.0001)
Months 7–9	.77	.09	.06	-.47	.0008
				(.09)	(.0001)
Months 10–12	.91	.06	.05	-.47	.0005
				(.09)	(.0001)
Months 13–18	.91	.06	.05	-.38	-.0002
				(.06)	(.0001)
Months 19–24	.92	.05	.04	-.42	.0002
				(.06)	(.0001)

<sup>a</sup> From regressions of log dependent variables on distributed lag of  $IPI^*$  (deviations from trend in the log of the Industrial Production Index), time and month-specific dummy variables. Standard errors are shown in parentheses. Elasticities reflect the sum of distributed lag coefficients.  $N=180$ .

<sup>b</sup> Shown in weeks.

<sup>c</sup> Shown in thousands.

<sup>d</sup> Percentage of labor force.

Table 4). A 10 percent decline in  $IPI^*$  increases continuation rates by 8 to 9 percent for individuals in their first six months of unemployment, about 5 percent for individuals in their second six months, and 4 percent for individuals in their second year.

Cyclical patterns in expected duration and incidence are also reexamined using monthly data in regressions that include  $IPI^*$ , a trend term, and month-specific dummy variables. Not surprisingly, the results (see Table 4) are very similar to those based on annual data

reported in Table 2. The cyclical elasticity of expected duration shows the equilibrium change in duration that would result from a steady-state change in  $IPI^*$ . The results show a cyclical elasticity of  $-3.2$  for the duration measure and  $-.60$  for incidence. The cyclical elasticity of the incidence rate is estimated to be  $-.69$ .

All measures follow a highly significant positive trend. The coefficient on the trend variable indicates that mean completed spell duration has increased about 3.4 percent per



year, *ceteris paribus*. The incidence rate has increased by roughly 1.8 percent per year while total incidence has increased by about 4.2 percent annually.

Recall that in a steady-state environment, unemployment is the simple product of incidence and mean duration. In turn, the sum of the long-run percentage responses of incidence and duration to business cycle conditions yields the long-run percentage response in total unemployment. The elasticities reported in Table 4 thus imply that a 1 percent (steady-state) decline in  $IPI^*$  yields an equilibrium increase in unemployment of roughly 3.8 percent (3.9 percent in the unemployment rate). In contrast, the sum of the elasticities from the steady-state model (reported in Table 2) indicate that a 1 percent decline in  $IPI^*$  would increase unemployment by only about 2.3 percent.

As a check on these estimates of the long-run effect of business conditions on the components of unemployment, regressions were estimated to examine directly cyclical fluctuations in total unemployment and the unemployment rate. The regressions included a distributed lag of  $IPI^*$ , a linear trend, and month-specific dummy variables. The results indicate an elasticity of the unemployment rate with respect to  $IPI^*$  of 3.0 percent and an elasticity of unemployment with respect to  $IPI^*$  of 2.8 percent. It is likely that these figures are somewhat smaller than the long-run impacts derived from the incidence and duration measures because recessions are generally too short for the larger entering cohorts to filter through the entire duration schedule. As a result, new (higher) steady-state unemployment levels are not achieved. The expected duration measure, recall, simulates an instantaneous adjustment to new economic conditions.

Of the resulting equilibrium response of unemployment to business conditions, fully 84 percent (3.2/3.8) results from changes in duration. The rest results from cyclical fluctuations in incidence. By means of contrast, the steady-state results reported by Perry attribute a more important role for changes in incidence for the years 1954 through 1971. For prime-age males, for

example, Perry attributes roughly 55 percent of changes in unemployment to changes in incidence. For females, the corresponding figure is roughly 60 percent.

In sum, the incidence and mean completed duration of unemployment spells are cyclical in nature. However, fluctuations in unemployment appear to result more from changes in spell duration and less from the changes in the incidence of new spells. The use of steady-state methods to derive such estimates dampens cyclical responses and understates the influence of duration on changes in unemployment.

## V. Conclusions and Qualifications

This paper examines cyclical changes in unemployment. These fluctuations can be described in terms of changes in the incidence and average duration of unemployment spells. As such, a primary focus of the paper is the development of unbiased measures of the incidence and duration of such events.

A new method for estimating mean completed spell length is developed and the results are compared with those based on the fairly restrictive assumption that unemployment reflects steady-state conditions. The observed biases introduced by steady-state techniques are consistent with those expected. Analysis of nonsteady-state measures indicates that duration and incidence are quite cyclical and that changes in duration play a predominant role in explaining both cyclical fluctuations and secular trends in unemployment.

These results are subject to some important qualifications. First, the underlying data do not permit distinguishing between spells which end due to (i) employment and (ii) departure from the labor force. Spells are not differentiated by reason for unemployment (entrant, layoff, job leaver, etc.). Estimates pertain to single spells as opposed to average unemployment experience (which may include multiple spells). Finally, the data reveal distinct patterns of response bias which necessitates the use of somewhat *ad hoc* smoothing techniques. The results, however,

seem fairly robust with respect to the choice of smoothing algorithm.

Despite these limitations, the estimates reveal some intriguing patterns in the composition of fluctuations and trends in aggregate unemployment, as officially defined. The determinants of large and significant positive trends in incidence and duration remain important and unanswered policy questions.

## REFERENCES

- Akerlof, George and Main, Brian, "Unemployment Spells and Unemployment Experience," *American Economic Review*, December 1980, 70, 885-93.
- \_\_\_\_\_ and \_\_\_\_\_, "An Experience-Weighted Measure of Employment and Unemployment Duration," *American Economic Review*, December 1981, 71, 1003-12.
- Antos, Joseph, Mellow, Wesley and Triplett, Jack, "What is a Current Equivalent of Unemployment Rates of the Past?," *Monthly Labor Review*, March 1979, 102, 36-46.
- Bowers, J. K. and Harkess, D., "Duration of Unemployment by Age and Sex," *Economica*, December 1979, 46, 239-60.
- Bowers, Norman, "Probing the Issues of Unemployment Duration," *Monthly Labor Review*, July 1980, 103, 23-32.
- \_\_\_\_\_ and Horvath, Francis, "Keeping Time: An Analysis of Errors in the Measurement of Unemployment Duration," *Journal of Business and Economic Statistics*, 1985 forthcoming.
- Carlson, John and Horrigan, Michael, "Measures of Unemployment as Guides to Research and Policy: Comment," *American Economic Review*, December 1983, 73, 1143-52.
- Clark, Kim and Summers, Lawrence, "Labor Market Dynamics and Unemployment: A Reconsideration," *Brookings Papers on Economic Activity*, 1:1979, 13-72.
- \_\_\_\_\_ and \_\_\_\_\_, "Demographic Differences in Cyclical Employment Variation," *Journal of Human Resources*, Winter 1981, 16, 61-77.
- Kaiz, Hyman, "Analyzing the Length of Spells of Unemployment," *Monthly Labor Review*, November 1970, 93, 11-20.
- Kiefer, Nicholas, Lundberg, Shelly and Neumann, George, "How Long is a Spell of Unemployment?: Illusions and Biases in the Use of CPS Data," mimeo., April 1983.
- Luckett, James P., "A Communication: Estimating Unemployment Duration," *Brookings Papers on Economic Activity*, 2:1979, 477-79.
- Perry, George L., "Unemployment Flows in the U.S. Labor Market," *Brookings Papers on Economic Activity*, 3:1972, 245-78.
- Poterba, James and Summers, Lawrence, "Survey Response Variation in the Current Population Survey," Working Paper No. 1109, National Bureau of Economic Research, April 1983.
- Salant, Stephen, "Search Theory and Duration Data: A Theory of Sorts," *Quarterly Journal of Economics*, February 1977, 91, 39-57.
- Tella, Alfred, "Cyclical Behavior of Bias-Adjusted Unemployment," in *Methods for Manpower Analysis*, No. 11, W. E. Upjohn Institute for Employment Research, April 1976.

# The Advantages of Being First

By A. GLAZER\*

In analyzing issues of entry into a market, different economists have chosen to emphasize different aspects of the issues. Some look at the advantages, such as economies of scale, learning by doing, or strategic opportunities, that early entrants possess over later ones (see for example, Avinash Dixit, 1979, 1980; B. Curtis Eaton and Richard Lipsey, 1979; Richard Schmalensee, 1982; and A. Michael Spence, 1979, 1981).

Other economists suppose that in equilibrium a firm should gain no advantage from entering before others. Richard Jensen (1982) uses this assumption to analyze the diffusion rate of an innovation among firms that learn over time and that have different prior beliefs about the innovation's profitability. Jennifer Reinganum (1981) describes the Nash equilibrium timing of entry in a world consisting of identical firms, but in which entry at different times entails different costs. J. Luis Guasch and Andrew Weiss (1980) examine the effect of uncertainty about workers' qualifications on the entry decisions of firms, and Takeo Nakao (1980) looks at entry under conditions of oligopoly.

Entry, however, should not be studied in isolation from failure. Alvin Star and Michael Massel (1981), for example, find that out of 17,252 retail establishments established in Illinois during the 1960's, only 33.2 percent survived after five years. J. Hugh Davidson (1976) finds that about 70 percent of test market brands were not expanded nationally, and could therefore be considered failures. Any model of entry into an industry must allow for the possibility of failure, and Boyan Jovanovic (1982) and Steven Lippman and R. P. Rumelt (1982) indeed do so. They, however, give little attention to the sequential process of firm entry.

We thus find two differing streams of thought: one that emphasizes the advantages of early entry, and one that focuses on the condition that firms earn identical profits. Integrating these views is a primary purpose of the theoretical part of this paper. The paper also complements the aforementioned studies by considering the possibilities of both entry and failure, and by treating market conditions as a random walk, which allows for a deeper analysis of the characteristics of innovation. The assumptions of the model are the topic of Section II, Section III describes the equilibrium entry times of firms, and Section IV explains why the failure rates of first and later entrants may be similar, and offers some results of relevance to government policy promoting innovation. It is most important, however, to first examine some empirical evidence concerning whether or not first entrants enjoy special advantages.

## I. Some Data

Several investigators find that the most successful existing firms are those that devoted the most resources to research and development, or that innovated the earliest (see Edwin Mansfield, 1962; Ira Whitten 1979). But the sample of existing firms is a biased one. It consists of *successful* innovators; little record is left of those innovators who failed. It could well be that in successful markets first entrants perform exceedingly well, but that any firm contemplating entry could do as well by entering an established market as by developing a new one in which the risk of collapse is high. Indeed, R. G. Cooper (1979), who studies 100 new product successes and failures, finds that being first in a market offered no particular advantage. William Dillon et al. (1979) reach a similar conclusion in their investigation of 109 firms that had introduced new products.

These studies, interesting though they are, rely on analyses of questionnaires sent to

\*School of Social Sciences, University of California, Irvine, CA 92717. I am grateful to an anonymous referee for helpful comments.

managers. It would be useful to examine the actual behavior of firms. Unfortunately, requisite data for most industries are not available, and a study must therefore be limited to an industry for which data can be found on the dates of entry and exit of all firms, including firms which failed soon after they were established. Perhaps the best such information is available for newspapers. In particular, *A Bibliography of Iowa Newspapers, 1836-1976* (see Alan Schroder, 1979) lists all daily newspapers published in that state; it also lists changes in the name of a newspaper, the date at which a newspaper began publication, and the date (if any) at which the newspaper merged with some other newspaper. These data thus permit a careful study of entry and exit in several dozen markets.

The first question I investigate using these data is whether first entrants in successful markets do possess some advantage over second entrants. In eighteen of the markets studied, a first and second entrant existed simultaneously for some length of time; presumably these were in some sense successful markets, markets attractive enough to attract at least two firms. Following the second entrant's appearance, the first entrant survived longer than did the second entrant in thirteen of these markets; such a skewed distribution could occur by chance with a probability of less than 5 percent. It appears that in a competition between first and second entrants, the first entrant possesses a significant advantage.

The results differ if one looks at all markets, rather than just at successful ones. In general, first entrants enter a market before second entrants do, and a statistical method must be devised to separate the effects of the date of entry from the order of entry. I therefore constructed a sample of newspapers of which half were first entrants and half were second entrants. To each first entrant in the sample founded in year  $t$ , there corresponds a second entrant (not necessarily in the same city) that was founded within five years of year  $t$ . The distribution of entry dates for first and second entrants studied is therefore similar. In this sample, which includes some markets in which only

one newspaper ever appeared, I found no significant difference between the survival rates of first and second entrants: twenty-five years after founding, first entrants had a survival rate of 32 percent, and second entrants a survival rate of 39 percent. The *chi-square* test for the difference in survival rates for the years 1, 2, 3, ..., 40 is significant at only the 35 percent level. First entrants did not appear to be more successful than later entrants.

In short, in successful markets first entrants did better than second entrants; but when one looks at all markets this superiority disappears. Explaining these results is the topic of the following sections.

## II. Assumptions

To highlight the essential features of the problem, and to make the analysis more tractable, I assume that there are only two potential entrants in the market—entrepreneurs  $A$  and  $B$ . Each of these entrepreneurs can enter the market at any time he wishes; entry entails a nonrecoverable fixed cost of  $F$  dollars. An entrepreneur can also exit the industry whenever he wishes, and he can later reenter the industry after again incurring a cost of  $F$  dollars.

The profits earned by a firm are a function of the firm's costs, of the market demand curve, of whether or not another firm is already established at the time the firm in question enters the market, and of the length of time that it and its competitors have been in business. For simplicity, suppose that the cost function is fixed over time, but that the market demand curve can shift over time. The characteristics of the demand curve can be summarized by the value of a parameter,  $s$ , which can represent, for example, the size of a city in which a product is sold, or the average income level of consumers. The firm's profits are an increasing function of  $s$ , other things held constant.

The value of  $s$  varies over time in a stochastic manner; more specifically, it follows a random walk. Let the passage of time be measured at discrete intervals of length  $\Delta t$ . Over each such interval  $s$  can increase or decrease by the fixed amount  $\Delta s$ . The prob-

ability of such an increase is  $p$ , and the probability of such a decrease is  $1-p$ . The value of  $p$  is constant over time, so that the stochastic process is stationary; that is, the probability that  $s$  takes on some value at time  $t + \Delta t$  depends only on the value of  $s$  at time  $t$ . It will be convenient to allow for growth in the market, which is translated into the assumption that  $p$  is greater than one-half.

The model I use considers discrete values of  $\Delta s$  and  $\Delta t$ . It will occasionally prove useful to examine a model in which the changes are very small and occur in very rapid succession, or where the limits of  $\Delta s$  and  $\Delta t$  are zero. Strictly speaking, the passage to the limit entails the study of diffusion processes. William Feller (1957, pp. 323-27) shows, however, that the equations for a random walk hold when the passage to such a limit is taken.

Suppose that at time zero the market is so small as to make entry then unprofitable. Over time the market is expected to grow, although the date at which  $s$  will attain some particular value is a random variable. In any one market, however, the random walk describing changes in the value of  $s$  will take some unique path, and it is therefore meaningful to speak of the first instant at which  $s$  attains a specified value. Define  $t(s^*)$  as the time to first passage, or the first instant at which  $s = s^*$ ; the value of  $s$  is less than  $s^*$  prior to time  $t(s^*)$ , and its value may be less than, equal to, or greater than  $s^*$  after time  $t(s^*)$ . Two possible paths for a random walk are shown in Figure 1 by the lines  $OA$  and  $OB$ . When the random walk takes path  $OA$ ,  $t(1)$  equals 1; when the path is  $OB$ ,  $t(1)$  equals 9.

Most generally, each firm's profits will depend not only on the state of the market, but also on the length of time it has been in business. We might expect a first entrant's profits following the entry of a second firm to be higher the longer the first firm was the sole one in the industry; the opposite relation might hold for the later entrant. Unfortunately, modeling such a general model presents great difficulties. A second firm would enter the market when the value of market demand became greater than some critical

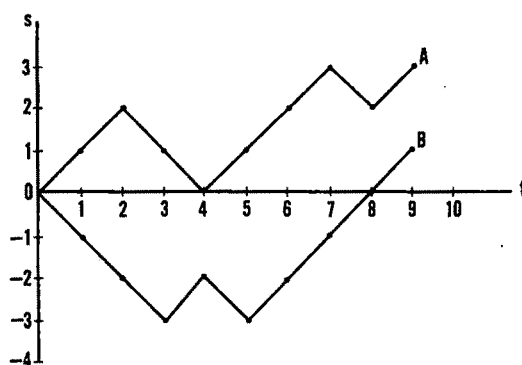


FIGURE 1

value, which value depends on the length of time that has elapsed since the other firm entered the market. Thus, a firm might enter the market when  $s = \bar{s}$  if the other firm was established one year ago, but would not enter when  $s = \bar{s}$  if the other firm was established three years ago.

I simplify matters by considering two possibilities only. The first is that a first entrant enjoys no advantages over a later entrant. The second possibility is that the advantages of early entry are totally independent of the length of time that elapses between the appearances of the first and the second firm.

Under these assumptions the firm's profit rate at time  $t$  depends on the value of  $s$  at that time, and, perhaps, on whether or not another firm preceded it. Given the value of  $s$  at any instant, an entrepreneur can calculate the probability distribution functions of  $s$  for all subsequent instants. The expected value of a firm's future profits, discounted to time  $t$ , is therefore a function only of the value of  $s$  at that time and of the number of competitors then established.

Denote by  $V^k(s)$  the expected value of a firm's future profits, discounted to time  $t(s)$ , given that it is the  $k$ th entrant into the industry; at the time of its entry  $k-1$  firms had already been established and were still in business. Let  $V^{1.5}(s)$  be a firm's expected future profits, discounted to time  $t(s)$ , if both firms simultaneously entered the industry at time  $t(s)$ .

I assume that profits are an increasing function of  $s$ , so that  $V^k(s + \Delta s) - V^k(s) > 0$ .

for all  $\Delta s > 0$ . For given values of  $s$  and of the number of competitors, an early entrant enjoys profits no lower than those enjoyed by later entrants:  $V^k(s)$  is a nonincreasing function of  $k$ .

Note that in calculating  $V^k(s)$  the firm takes into account the possibility that  $s$  may become so low that it would leave the industry. This is especially likely to occur if the firm enters the market at a time when  $s$  is already low. An entrepreneur can also consider the possibility that at some future date another firm will enter the market. In short, the value of  $V^k(s)$  incorporates all factors that affect a firm's profits.

It will prove useful to have a notation for a firm's expected profit during finite periods of time. Let  $\pi^k(s_1, s_2)$  represent the firm's expected profits during the interval  $[t(s_1), t(s_2)]$ , discounted to time  $t(s_1)$ , given that the firm was the  $k$ th entrant into the industry, and that during the interval  $[t(s_1), t(s_2)]$  no new firms entered. Recall that  $t(s)$  is a random variable, and that from the instant at which  $s = s_1$ , until the instant at which  $s = s_2$ , the value of  $s$  is a random variable as well (compare in Figure 1 the paths  $OA$  and  $OB$  which both start at  $t(0) = 0$  but which then follow different paths). Thus, the actual profits earned by the firm during any interval of time is a random variable, although the firm knows the expected value of those profits.

Finally, define  $R(s_1, s_2)$  as the expected discount factor. More specifically, for  $s_2 > s_1$  let  $f(t, s_1, s_2)$  be the probability that  $t(s_2) - t(s_1) = t$ , let  $r$  be the discount rate, and let

$$R(s_1, s_2) \equiv \sum_{t=0}^{\infty} f(t, s_1, s_2)(1+r)^{-t}.$$

### III. Equilibrium Entry Times

The analysis is an interesting one only if there will be at least two entrants which do not enter simultaneously. I shall present some assumptions that ensure these results.

Assume that for some sufficiently high value of  $s$ , say  $\hat{s}$ , the value of  $V^2(\hat{s})$  is greater than  $F$ . A standard result in the theory of random walks among an infinite number of states (E. B. Dynkin and V. A.

Uspenskii, 1963, p. 51) is that if the transition probabilities from  $s$  to  $s + \Delta s$  are all greater than or equal to  $1/2$ , then with probability 1 each state will be reached within a finite number of periods. The second firm will enter when, say,  $s = s^* \geq \hat{s}$ , and under the just-stated conditions this will occur with probability one. The first firm will enter no later than at time  $t(s^*)$ .

Consider next the question of whether entry will occur simultaneously. Define  $t(s_A)$  as firm  $A$ 's entry time, and let firm  $B$  correctly anticipate this entry. Suppose for the moment that firm  $A$  is not the latest entrant. Entrepreneur  $B$ 's problem is then to choose a value of  $s$ , where  $s \geq s_A$ , and thus an instant  $t(s) \geq t(s_A)$ , such that entry at that instant would maximize his profits. Entrepreneur  $B$ 's profits, discounted to time  $t(s_A)$ , are then:

for entry at time  $t(s_A)$ ,

$$V^{1.5}(s_A) - F = \pi^{1.5}(s_A, s) - F \\ + V^{1.5}(s)R(s_A, s);$$

for entry later than at  $t(s_A)$ ,

$$(V^2(s) - F)R(s_A, s).$$

A necessary condition for  $t(s_A)$  to be firm  $B$ 's optimal entry time is that

$$(1) \quad \pi^{1.5}(s_A, s_A + \Delta s) \geq [V^2(s_A + \Delta s) \\ - V^{1.5}(s_A + \Delta s)]R(s_A, s_A + \Delta s) \\ + F[1 - R(s_A, s_A + \Delta s)]$$

for  $\Delta s > 0$ .

Analogously, we can examine the profitability of firm  $B$  entering the industry prior to time  $t(s_A)$ . A necessary condition for  $t(s_A)$  to be firm  $B$ 's optimal time is that

$$(2) \quad \pi^1(s_A - \Delta s, s_A) \leq [V^{1.5}(s_A) - V^1(s_A)] \\ \times R(s_A - \Delta s, s_A) + F[1 - R(s_A - \Delta s, s_A)].$$

To recapitulate, each firm will wish to enter at time  $t(s_A)$  only if doing so maximizes its expected profits given that the other firm

enters then, or, equivalently, only if both inequalities (1) and (2) hold.

Suppose that the functions  $\pi^k$  and  $V^k$  are continuous in  $\Delta s$ , that  $\lim_{\Delta s \rightarrow 0} R(s, s + \Delta s) = 1$ , and that  $\lim_{\Delta s \rightarrow 0} \pi^k(s, s + \Delta s) = \lim_{\Delta s \rightarrow 0} \pi^k(s - \Delta s, s) = 0$ . (What this means is that very small changes in  $s$  occur in arbitrarily short periods of time.)<sup>1</sup>

Consider first the possibility that the advantages of first entry are noninfinitesimal even if the second firm enters immediately after the first. That is,  $[V^{1.5}(s_A) - V^1(s_A)] \neq 0$ , but is instead a negative number. In that case, inequality (2) cannot hold for all  $\Delta s$ , and entry will not occur simultaneously.

Consider next the possibility that there are no advantages to early entry: the value of  $V^2(s_A + \Delta s) - V^{1.5}(s_A + \Delta s)$ , which appears in inequality (1), is zero, and the value of  $V^{1.5}(s_A) - V^1(s_A)$  which appears in equality (2), is also zero. Observe that the value of  $R(s_1, s_2)$  depends only on the value of  $s_2 - s_1$ , and not on the levels of  $s_1$  or  $s_2$ ; thus  $R(s_A, s_A + \Delta s)$  appearing in expression (1) is equal to  $R(s_A - \Delta s, s_A)$  appearing in expression (2).

In this case, simultaneous entry will occur only if  $\lim_{\Delta s \rightarrow 0} \pi^{1.5}(s_A, s_A + \Delta s) \geq \lim_{\Delta s \rightarrow 0} \pi^1(s_A - \Delta s, s_A)$ . Note that, in general,  $\pi^1(s, s + \Delta s) > \pi^{1.5}(s, s + \Delta s)$  because  $\pi^1(\cdot)$  represents expected profits earned by a monopolist during some interval, whereas  $\pi^{1.5}(\cdot)$  represents the profits earned by each duopolist. Let  $\pi^1(s, s + \Delta s) - \pi^{1.5}(s, s + \Delta s)$  equal  $c\Delta s$ , where  $c$  is some nonnegative number. Then simultaneous entry occurs only if  $d\pi/ds > c$ , or only if a firm's profits are affected more by a slight change in market conditions than by the entry of an additional firm. This possibility is not impossible, but appears to be quite unlikely. The result can be summarized as follows.

**PROPOSITION 1:** *Sufficient conditions for nonsimultaneous entry are that either of the*

*following conditions hold:*

- (a)  $V^1(s) - V^{1.5}(s) > 0$ ;
- (b)  $\lim_{\Delta s \rightarrow 0} [\pi^1(s, s + \Delta s) - \pi^{1.5}(s, s + \Delta s)]$   
 $> \lim_{\Delta s \rightarrow 0} [\pi^1(s, s + \Delta s) - \pi^1(s - \Delta s, s)]$ .

I shall henceforth assume that condition (b) holds for all  $s$ . I shall allow for the possibility that (a) does not hold, or that  $V^k(s) = V^j(s)$  for all  $k$  and  $j$ .

**PROPOSITION 2:** *Consider a Nash equilibrium in which each entrepreneur correctly anticipates the entry time of the other firm. In such an equilibrium the two firms earn identical profits.*

**PROOF:**

Consider an equilibrium in which firm  $A$  enters first, at time  $t(s_A)$ . Firm  $B$ , the second entrant, will enter when  $s$  first attains a value  $s^*$ , defined as the lowest value greater than  $s_A$  for which

$$(3) \quad \pi^2(s^*, s^* + \Delta s) \geq FR(s^*, s^* + \Delta s),$$

$$(4) \quad \pi^2(s^* - \Delta s, s^*) \leq FR(s^* - \Delta s, s^*).$$

Firm  $B$ 's expected profits, discounted to time 0, are

$$Z_B(s^*) = [V^2(s^*) - F] R(0, s^*).$$

Firm  $A$ , the first entrant, has expected profits of

$$Z_A(s_A, s^*) = [\pi^1(s_A, s^*) - F] R(0, s_A) \\ + V^1(s^*) R(0, s^*).$$

Suppose first that  $Z_A$  is greater than  $Z_B$ . Then firm  $B$  could increase its profits by adopting the following strategy. Instead of entering when  $s = s^*$ , it enters when  $s = s_A - \Delta s$ , for arbitrarily small  $\Delta s$ . Once such entry occurs, firm  $A$  will enter when  $s = s^*$ . With the continuity of  $\pi(s, s + \Delta s)$ ,

<sup>1</sup>For any meaningful process,  $\lim_{\Delta s \rightarrow 0} [t(s + \Delta s) - t(s)] = 0$ ; otherwise  $t(s_2) - t(s_1)$  would have non-finite values for  $s_2 > s_1$ . This is why  $\lim_{\Delta s \rightarrow 0} \pi^k(s, s + \Delta s) = 0$ .

firm  $B$ 's profits will then be  $\lim_{\Delta s \rightarrow 0} Z_A(s_A - \Delta s, s^*) = Z_A(s_A, s^*)$ , which by assumption are greater than  $Z_B(s^*)$ . Thus, entry by the two firms with the first entrant earning larger profits cannot be an equilibrium.

Consider next a putative equilibrium in which the second entrant earns larger profits; that is,  $Z_B(s^*) > Z_A(s_A, s^*)$ . Then firm  $A$  can increase its profits by entering at time  $s^* - \Delta s$  for arbitrarily small  $\Delta s$ . From inequality (3),  $\pi^2(s^*, s^* + \Delta s) \geq FR(s^*, s^* + \Delta s)$ . Proposition 1 implies that  $\lim_{\Delta s \rightarrow 0} [\pi^1(s^* - \Delta s, s^*) - \pi^2(s^*, s^* + \Delta s)] > 0$  and therefore that  $\pi^1(s^* - \Delta s, s^*) > FR(s^* - \Delta s, s^*)$ . This implies that

$$\begin{aligned} Z_A(s^* - \Delta s, s^*) &> Z_A(s^*, s^*) \\ &\equiv [V^{1.5}(s^*) - F] R(0, s^*) \\ &\geq [V^2(s^*) - F] R(0, s^*) \\ &= Z_B(s^*). \end{aligned}$$

Thus, a situation under which  $Z_A(s_A, s^*)$  is less than  $Z_B(s^*)$  cannot be an equilibrium one. This completes the proof of Proposition 2.

**PROPOSITION 3:** *In equilibrium the first firm will enter the market at a time earlier than that which would maximize its profits in the absence of the threat of entry by another firm.*

**PROOF:**

In discussing Proposition 2, I showed that  $Z_A(s^* - \Delta s, s^*)$  is greater than  $Z_B(s^*)$  for some arbitrarily small  $\Delta s$ . Now for some sufficiently low levels of  $s$ ,  $Z_A(s, s^*)$  is negative. Assuming continuity of  $Z_A(s, s^*)$ , there therefore exists a value of  $s_A < s^* - \Delta s$  such that  $Z_A(s_A, s^*) = Z_B(s^*) < Z_A(s^* - \Delta s, s^*)$ . In equilibrium, firm  $A$  enters when  $s = s_A$ . Were firm  $A$  the only possible entrant, it could earn larger profits by entering at a later time, such as at time  $t(s^* - \Delta s)$ . Proposition 3 is thereby proved.<sup>2</sup>

<sup>2</sup>Reinganum obtained similar results in her model which considers firm adjustment costs rather than changing market conditions. Eaton and Lipsey show

#### IV. The Incentives to Innovate

The previous section showed that during the interval  $[t(s_A), t(s^*)]$  the first entrant enjoys a monopoly. This does not mean, however, that the second entrant will appear immediately after the first one does. Thus, the view (see Morton Kamien and Nancy Schwartz, 1982, p. 124) that perfect competition implies *immediate* imitation is not generally valid. This observation in turn leads to

**PROPOSITION 4:** *An innovator can earn economic profits greater than its competitor even if the latter can costlessly and immediately imitate an innovation once it appears on the market.*

**PROOF:**

Suppose an entrepreneur invents a new product. If all firms knew of this innovation, the entrepreneur would be forced to enter the industry at time  $t(s_A)$ , and earn profits no greater than those earned by the imitator who enters at time  $t(s^*)$ . Given secrecy, however, so that the product is concealed until the actual time of entry, the innovator could delay his entry until time  $s^* - \Delta s$  and earn profits of  $Z_A(s^* - \Delta s, s^*)$  which are greater than the profits of  $Z_B(s_A^*) = Z_A(s_A, s^*)$  that will be earned by the imitator. (Of course, entry when  $s = s^* - \Delta s$  need not be the innovator's optimal entry time.) That is, by inventing a new product and keeping its invention secret until it enters the market, the innovating firm can enter *later* than it would have to under conditions of perfect information, and can thereby earn supra-normal profits.

---

that in a spatial model the existing firm will expand rather than allow new firms to enter the industry; in a market for a homogeneous good as described here, this need not occur. Similarly, Ronald Johnson and Allen Parkman (1983) argue that in a market with room for only one firm, the first entrant will appear at the moment expected future profits become zero. They theorize that the entrant's profits should show an increasing trend over time, and find only weak evidence for that. They do not, however, consider the possibility of firm *failure*, and thus, I believe, give an incomplete test of their hypothesis.



Once the innovation is introduced, other firms can immediately copy it and enter the industry. As seen in the previous section, however, the second firm will not find it profitable to enter immediately, so that the innovating firm will enjoy monopoly profits for some length of time. Clearly the innovator would earn even greater profits if no firm could imitate the product: imitation reduces, and need not eliminate, the innovator's profits.

One would expect that an increase in the profits that a firm can earn would cause entry to occur earlier. This will be true, however, only in special cases.

**PROPOSITION 5:** *Let  $\pi^k(S)$  be a firm's rate of profit during the interval  $(t, t + \Delta t)$  given that  $k$  firms are in the industry and given that  $s = S$  at time  $t$ . An exogenous increase in  $\pi^k(S)$  need not hasten the date of first entry.*

**PROOF:**

Suppose that  $\pi^k(S)$  is exogenously increased for all values of  $s$  greater than  $s^*$ , where  $s^*$  is the value of  $s$  at which the second firm will enter the industry.<sup>3</sup> In the initial equilibrium the first entrant's expected profits, discounted to time zero, are

$$\begin{aligned} Z_A = & Z_B(s^*) + \pi^1(s_A, s^*) R(0, s_A) \\ & - F [R(0, s_A) - R(0, s^*)] \\ & + [V^1(s^*) - V^2(s^*)] R(0, s^*). \end{aligned}$$

The described increase in  $\pi^k(S)$  does increase  $Z_B$ , but it has no effect on  $\pi^1(s_A, s^*)$ , on  $F$ , or on  $R$ .

The effect of such a change on the value of  $V^1(s^*) - V^2(s^*)$  is ambiguous. The increase in  $\pi^k(S)$  for all  $s > s^*$  may increase the advantages of early entry, so that  $V^1(s^*) - V^2(s^*)$  increases. In equilibrium  $Z_A$  must equal  $Z_B(s^*)$ , but with the new value of  $V^1(s^*) - V^2(s^*)$ ,  $Z_A$  would be greater than

$Z_B$ . A new Nash equilibrium would require that firm  $A$  enter earlier than before, thereby reestablishing the equality of  $Z_A$  and  $Z_B(s^*)$ . It should be clear that if  $V^1(s^*) - V^2(s^*)$  decreased, first entry would be delayed; if changes in  $\pi^k(S)$  caused no changes in  $V^1(s^*) - V^2(s^*)$  the timing of entry would also be unchanged. In short, an improvement in expected profitability might delay rather than hasten entry. A similar result holds for changes in  $\pi(S)$  for values of  $s < s^*$ , and for changes in the transition probabilities from one value of  $s$  to another.

Finally, I can solve the puzzle that was a major theme of this paper. Recall that the empirical evidence showed that in markets consisting of two or more firms, the first entrant was more successful than the second. Yet first entrants in general were as likely to fail as second entrants. That is, it appears that first and later entrants earn identical expected profits even though early entrants enjoy strategic or other advantages. My model suggests that competition forces the first entrant to appear at an early stage in the development of a market so that there is a great danger that demand will not increase to the extent predicted. Therefore the market will not reach the necessary critical size in a reasonable time, and the first entrant will fail even before any other firms enter the industry. Later entrants will choose to enter only successful markets which have reached an appreciable size; their problem is the stiff competition given by earlier entrants. Observers who look only at the performance of early entrants in successful markets will overestimate the advantages of innovation.

## REFERENCES

- Cooper, R. G., "The Dimensions of New Product Success and Failure," *Journal of Marketing*, Summer 1979, 43, 93-103.
- Davidson, J. Hugh, "Why Most New Consumer Brands Fail," *Harvard Business Review*, March/April 1976, 54, 117-22.
- Dillon, William R., Calantore, Roger and Worthing, Parker, "The New Product Problem: An Approach For Investigating Product Failures," *Management Science*, December 1979, 25, 1184-96.

<sup>3</sup> Recall that although the second entrant enters when  $s = s^*$ , the random walk process may cause  $s$  to fall to a level below  $s^*$  after such entry occurs.

- Dixit, Avinash, "A Model of Duopoly Suggesting a Theory of Entry Barriers," *Bell Journal of Economics*, Spring 1979, 10, 20-32.
- \_\_\_\_\_, "The Role of Investment in Entry Deterrence," *Economic Journal*, March 1980, 90, 95-106.
- Dynkin, E. B. and Uspenskii, V. A., *Random Walks*, Chicago: University of Chicago Press, 1963.
- Eaton, B. Curtis and Lipsey, Richard G., "The Theory of Market Pre-emption: The Persistence of Excess Capacity and Monopoly in Growing Spatial Markets," *Economica*, May 1979, 46, 149-58.
- Feller, William, *An Introduction to Probability Theory and its Applications*, Vol. I, 2d ed., New York: Wiley & Sons, 1957.
- Guasch, J. Luis and Weiss, Andrew, "Adverse Selection by Markets and the Advantage of Being Late," *Quarterly Journal of Economics*, May 1980, 94, 453-66.
- Jensen, Richard, "Adoption and Diffusion of an Innovation of Uncertain Profitability," *Journal of Economic Theory*, June 1982, 27, 182-93.
- Johnson, Ronald and Parkman, Allen, "Spatial Monopoly, Non-Zero Profits and Entry Deterrence: The Case of Cement," *Review of Economics and Statistics*, August 1983, 65, 431-39.
- Jovanovic, Boyan, "Selection and the Evolution of Industry," *Econometrica*, May 1982, 50, 649-70.
- Kamien, Morton I. and Schwartz, Nancy L., *Market Structure and Innovation*, Cambridge: Cambridge University Press, 1982.
- Lippman, S. A. and Rumelt, R. P., "Uncertain Imitability: An Analysis of Interfirm Differences in Efficiency Under Competition," *Bell Journal of Economics*, Autumn 1982, 13, 418-38.
- Mansfield, Edwin, "Entry, Gibrat's Law, Innovation, and the Growth of Firms," *American Economic Review*, December 1962, 52, 1023-51.
- Nakao, Takeo, "Demand Growth, Profitability, and Entry," *Quarterly Journal of Economics*, March 1980, 94, 397-422.
- Reinganum, Jennifer F., "Market Structure and the Diffusion of New Technology," *Bell Journal of Economics*, Autumn 1981, 12, 618-24.
- Schmalensee, Richard, "Product Differentiation Advantages of Pioneering Brands," *American Economic Review*, June 1982, 72, 349-65.
- Schroder, Alan, *A Bibliography of Iowa Newspaper, 1836-1976*, Iowa City: Iowa State Historical Department, 1979.
- Spence, A. Michael, "The Learning Curve and Competition," *Bell Journal of Economics*, Spring 1981, 12, 49-70.
- \_\_\_\_\_, "Investment Strategy and Growth in a New Market," *Bell Journal of Economics*, Spring 1979, 10, 1-19.
- Star, Alvin D. and Massel, Michael Z., "Survival Rates for Retailers," *Journal of Retailing*, No. 2, 1981, 57, 87-99.
- Whitten, Ira Taylor, "Brand Performance in the Cigarette Industry and the Advantage to Early Entry, 1913-1974," Federal Trade Commission Staff Report, 1979.

# Internal Migration and Urban Employment in the Third World

By WILLIAM E. COLE AND RICHARD D. SANDERS\*

For more than a decade Michael Todaro's model has provided a widely accepted theoretical framework for explaining the massive flows of rural urban migration that are observed in many Third World countries.<sup>1</sup> This paper, however, presents data amassed from a wide range of countries to highlight a crucial shortcoming of that model. It will be shown that far from being general in nature, the Todaro approach is limited to explaining the movement of persons possessed of sufficient human capital to qualify them for modern sector employment. Masses of relatively uneducated persons migrate and work in a subsistence world that cannot be explained by the structures of Todaran theory. In turn, the present work utilizes the perspective of the urban subsistence sector to develop a model that serves as a useful complement to that of Todaro.

## I

The burden of the Todaro model was to explain why masses of workers moved from the countryside to the city in the face of sizeable urban pools of unemployed and underemployed. To accomplish this, the model focused attention on the present value of expected earnings rather than current wage rates. Specifically, the rate of rural-urban migration was held to be a function of the difference between the present values of expected urban earnings and expected rural earnings, with the size of the flow of expected urban earnings significantly affected by the probability of obtaining employment

in the urban modern (*U-M*) sector (Todaro, 1979, p. 201).<sup>2</sup>

Even though there might exist an urban pool of underemployed and unemployed labor, a potential migrant would decide to make the cityward trek if the expected *U-M* earnings, properly discounted by the probability factor, exceeded the expected stream of rural earnings. Those migrants not obtaining *U-M* employment in the immediate period are said to accept temporary employment in the urban surplus labor pool (Todaro, 1969, pp. 142 ff.). Although sometimes referring to the surplus labor pool as the traditional sector, the Todaro model clearly assumes that all members of the pool, as well as all migrants, are intent upon eventual *U-M* sector employment.<sup>3</sup> The Todaro model thus explains why there may be more migrants than modern sector job openings and accounts for the growth of the urban pool of surplus labor.

## II. Empirical Evidence

One of the most significant structural changes recently occurring in the Third

<sup>2</sup>Todaro's basic behavioral equation can be shown as

$$V(0) = \int_{t=0}^{\infty} [P(t)Y_u(t) - Y_r(t)] e^{-rt} dt - C(0),$$

where  $V(0)$  is the discounted present value of the net gain from a rural-urban move;  $P(t)$  represents the probability of securing a job in the modern urban sector in period  $t$ ;  $Y_u$ ,  $Y_r$  represent average real income in the modern urban and rural sectors, respectively;  $C$  is the one-time cost of the move; and  $r$  is the migrant's time preference rate of discount. In cases where  $V(0)$  is positive, the economically rational potential migrant will decide to move. All of Todaro's work has assumed that open rural unemployment is nonexistent. A probability of unity is therefore used when calculating expected rural earnings.

<sup>3</sup>This assumption is imbedded in Todaro's treatment of the probability of employment in the *U-M* sector. Other factors constant, that probability varies inversely with the size of the surplus labor pool (see Section III).

\*Department of Economics and Department of Statistics, respectively, University of Tennessee, Knoxville, TN 37916. This paper has benefited from comments by Walter C. Neale and Henry Thompson. The research support of the College of Business Administration is greatly appreciated.

<sup>1</sup>The seminal work of Todaro appeared in 1969 and successive formulations have been faithful to that work.

World has been the rapid growth of an urban subsistence (*U-S*) sector. Elsewhere called the urban informal sector or the urban traditional sector,<sup>4</sup> it comprehends those urban employment categories that feature very low levels of productivity and earnings. The *U-S* sector includes such occupations as domestic service, petty tradesmen (including street vendors), artisans, and the like.

The relative importance of the *U-S* sector can be seen from data reflecting the economic structure of several important cities. It has been estimated that the proportion of Calcutta's urban labor force relegated to the *U-S* sector is 43 percent; for Bogota, 45 percent; and 50 percent in Lagos (Harold Lubell, 1978, p. 753).<sup>5</sup> Census data for Mexico indicate that on the order of 34 percent of the labor force of the huge Federal District may be classified in the *U-S* sector (Secretaría de Industria Y Comercio, 1972). It is widely noted that these sectors exist and are growing in many Third World cities, due in large measure to rural-urban migration.

Two distinguishing characteristics of *U-S* sector employment are very low capital-labor ratios and few if any formal human capital requirements. One therefore finds few if any barriers to entry into *U-S* sector employment. On the other hand, it is widely recognized that *U-M* sector jobs carry education requirements that effectively exclude persons who have acquired little or no formal education. It has been further noted that, for given *U-M* jobs, educational prerequisites tend to escalate as the average level of education rises.<sup>6</sup> (Lisa Peattie, 1968, pp. 137–39; Todaro, 1979, pp. 251 ff.)

<sup>4</sup>Our term "urban subsistence sector" is in large measure synonymous with the widely used term "urban informal sector." Other terms found in the literature are "urban traditional" (Todaro, 1969), "lower circuit" (Milton Santos, 1979), and "protoproletariat" (T. G. McGee, 1971). There are important differences, however, which are explored later when a formal definition of the *U-S* sector is introduced. (See fn. 18 below and the discussion to which it pertains.)

<sup>5</sup>Lubell's terminology is "urban informal sector."

<sup>6</sup>The phenomenon of escalating job prerequisites is sufficiently widespread to have earned the title "credentialism." Jorge Balán et al. noted: "Given a choice provided by an abundant supply of labor,... modern

Mexican census data reveal that the bulk of migration into the Metroplex comes from the most rural and relatively poor states where levels of education are absolutely and relatively low. (Although the present paper treats rural-urban migration in general terms, it utilizes as a reference point, the large urban agglomeration in and around the Federal District of Mexico, which we call the Metroplex.)<sup>7</sup> Surveys confirm that the majority of migrants have minimal educational attainments (Humberto Muñoz et al., 1977, p. 109, and Larissa Lomnitz, 1977, p. 88). There is also evidence that the average education attained by migrants coming into the Metroplex has been falling and that the percentage of migrants originating in strictly rural environs has been growing (Muñoz et al., pp. 106–07), an important distinction between current and earlier migration patterns.

An extensive survey found that over the span of several decades the mean level of education of urban immigrants had regressed toward the rural mean (Harley Browning and Waltraut Feindt, 1969). Migration is therefore becoming less and less selective. It appears then that many persons move to the city with the expectation of finding long-term employment in the *U-S* sector,<sup>8</sup> and that

---

enterprise in Latin America often set educational requirements (credentials) of some sort for all positions, even though the actual performance of tasks may require little formal education" (1973, p. 17). When writing on the topic of education, Todaro noted that scarce jobs are often allocated on the basis of ever increasing educational "credentials" (1979, pp. 250–51). Malcolm Gillis et al. refer to the process as "educational deepening" (1983, p. 221), and R. M. Sundrum allows that education used as a screening device has a perverse economic impact (1983, p. 89). Whatever the term, it is clear that education prerequisites constitute serious barriers to *U-M* sector employment.

<sup>7</sup>It should be noted that Todaro's work had its empirical base in East Africa, and especially Kenya.

<sup>8</sup>This is not in contradiction of previous studies that found a positive correlation between educational attainment and migration. The greater an individual's educational attainment, the less likely is rural employment that will compensate for the human capital investment. Therefore, the higher the average level of rural education, the greater will be the level of outmigration, *ceteris paribus*. Both data on education and data on migration

their decision is based upon information obtained through a system so thorough and intricate that anthropologists label it a network.<sup>9</sup> The Todaro model, on the other hand, posits that urban emigres are attracted by prospects of employment in the *U-M* sector. It therefore explains the movement of migrants who deem themselves to possess the qualifications necessary for such employment. It does not explain the movement of those intent upon *U-S* sector employment. Indeed, persons migrating in the face of zero probability of employment in the *U-M* sector must be seen as irrational by the Todaro model.

### III

In approaching Todaro's treatment of the *U-S* sector, we begin by investigating his specification of the probability of obtaining *U-M* employment. Specifically, the Todaro formulation for the probability of being selected from the urban "surplus labor" pool during any one time period  $\pi(t)$  is<sup>10,11</sup>

$$\pi(t) = \gamma N_u(t) / (N(t) - N_u(t)),$$

reflect dual phenomena, however. In rural environs, the average level of education grows as does the number of uneducated. Both educated and uneducated join the trek to the cities.

<sup>9</sup>This is described in detail in Lomnitz.

<sup>10</sup>This is parallel to equation (6) in Todaro's seminal work (1969).

<sup>11</sup>The probability term  $\pi$  is closely related to the probability,  $P$ , that is found in fn. 2. In any given time period, the probability of being employed in the *U-M* sector,  $P(t)$ , is related to the probability,  $\pi$ , of having obtained employment in that or any previous time period. This is shown in the following (Todaro, 1979, p. 202):

$$P(x) = \pi(1) + \sum_{t=2}^x \pi(t) \prod_{s=1}^{t-1} [1 - \pi(s)]$$

where  $x$  is the number of time periods and where employment selection is random so that  $\pi(t)$  equals the number of job openings in period  $t$  divided by the number of job aspirants in that period. The probability of obtaining modern sector employment is therefore seen as improving over time.

where  $\gamma$  is the rate of growth of employment in the *U-M* sector,<sup>12</sup>  $N$  represents the total urban labor force, and  $N_u$  is the employed *U-M* labor force.

Although the specification appears straightforward, Todaro displayed indecision in his interpretation of that urban labor component not employed in the modern sector,  $N - N_u$ . At one point, this term is explicitly held to represent "a large pool of unemployed and underemployed who arrived in town earlier and still are waiting for a modern sector job" (1969, p. 142). This labor pool is several times referred to as the "urban traditional sector"; more or less corresponding to what we label as the *U-S* sector. However, the term  $(N - N_u)/N$  is also formally defined as the unemployment rate.<sup>13</sup> With regard to the internal logic of the model, it may not matter which rubric attaches to that part of the urban labor force not employed in the *U-M* sector. In terms of empirical relevancy, however, it makes a world of difference. Rates of open unemployment tend to be relatively low while the *U-S* sector tends to be relatively large and growing. In the Metroplex, for example, open unemployment was around 4 percent at a time when the *U-S* sector accounted for about 34 percent of the urban labor force. At the same time, employment in the *U-M* sector of the Metroplex was growing at an annual rate of 3.6 percent. If the unemployment rate was used to calculate  $\pi$ , the probability of a migrant finding a *U-M* job in the subject year would have been .86. On the other hand, if the *U-S* sector was used as Todaro's  $N - N_u$ ,  $\pi$  would have been .084. In the former case, a migrant's prospects for *U-M* employment would have been quite good; in the latter case, they would have appeared dim.

<sup>12</sup>Where  $\lambda$  is the rate of growth of industrial output and  $\rho$  is the rate of labor productivity growth,  $\gamma = \lambda - \rho$ .

<sup>13</sup>See the discussion pertaining to equation (10) in Todaro (1969). Elsewhere, Todaro makes the straightforward statement that "the probability of obtaining an urban job is inversely related to the urban unemployment rate" (1976, p. 196). A number of empirical tests of the Todaro model have utilized open unemployment (Todaro, 1976, pp. 68 ff.).



TABLE 1—PROBABILITY ( $\pi$ ) OF IMMEDIATELY OBTAINING A *U-M* JOB

$\gamma$	$(N - N_u)/N$					
	.05	.10	.20	.30	.40	.50
.01	.19	.09	.04	.02	.02	.01
.02	.38	.18	.08	.05	.03	.02
.03	.57	.27	.12	.07	.05	.03
.04	.76	.36	.16	.09	.06	.04
.05	.95	.45	.20	.12	.08	.05
.06	1.14	.54	.24	.14	.09	.06
.07	1.33	.63	.28	.16	.11	.07

Note:  $\gamma = \lambda - \rho$ , when  $\lambda$  is the annual rate of growth of industrial output and  $\rho$  is the annual rate of growth of labor productivity in the *U-M* sector.

We will further assess the explanatory range of the Todaro model through exercises that assign a range of plausible values to several of Todaro's variables. In Table 1, several reasonable annual rates of employment creation,  $\gamma$ , are combined with a range of values for the relative size of urban surplus labor pool to produce possible values of  $\pi$ , the probability of obtaining a *U-M* job in the immediate period.

Because development includes growth of labor productivity, widely observed annual growth rates of from 5 to 10 percent for *U-M* sector output likely promote annual employment growth,  $\gamma$ , in the range of 2 to 7 percent. If the surplus labor pool from which *U-M* hiring takes place is defined as open unemployment, then values for  $(N - N_u)/N$  of .05 or .10 would probably be relevant. For those combinations of  $\gamma$  and  $(N - N_u)/N$ , the values for  $\pi$  are relatively high, but because rates of open unemployment are used, the model, in this instance, cannot explain the existence, size, or growth of the *U-S* sector. On the other hand, if the surplus labor pool is held to be the *U-S* sector, then values for  $(N - N_u)/N$  of 30 to 50 percent more likely apply. In those cases the more plausible values for  $\pi$  are observed to be quite low.

It will now be appropriate to relate  $\pi$  to expected income. In Table 2, we combine a range of representative values for  $\pi$  (taken from Table 1) with various possible rates of time-preference discount  $r$ , and determine the number of years that would be required before the present value of expected urban earnings would equal the present value of

TABLE 2—YEARS REQUIRED TO ELIMINATE PRESENT VALUE DIFFERENTIALS

$\pi$	$r$		
	.05	.10	.15
.03	> 50	> 50	> 50
.06	38	> 50	> 50
.09	20	33	> 50
.12	12	16	28
.15	10	11	13
.18	8	8	9

expected rural earnings in a situation where the *U-M* wage rate is twice that of the rural sector.<sup>14</sup> The values chosen for  $\pi$  imply that the *U-S* sector is the pool from which new hires are drawn for the *U-M* sector.<sup>15</sup>

<sup>14</sup>The solutions relate to the values of  $\pi$ , drawn from Table 1 and held constant in each of the time intervals, to selected values of  $r$ , the time-period-preference discount factor, and to a modern wage that is twice the rural wage. Present value calculations are based on the equation for  $V(0)$  found in fn. 2. The probability term  $P(t)$  found in that equation relates to  $\pi$  in the manner specified in fn. 11. As with Todaro, the rural employment probability is considered to be unity. Because the Todaro model calculates incomes in real terms, future earnings are converted to constant currency of the base period. The time-preference rate,  $r$ , used to determine the present values of future income streams should therefore be stated in market values for the base period,  $t(0)$ . If "real pesos" refer to purchasing power in  $t(0)$ , then a real peso earned at some future time might properly be discounted by a real interest rate.

<sup>15</sup>In Todaro's model, new hires are selected at random from the surplus labor pool,  $N - N_u$ . One can conceptualize a more elaborate Todaro model which utilizes various values for  $\pi$  based upon respective stocks of human capital of several groups of potential migrants.

TABLE 3—CHARACTERISTICS OF THE BASIC SECTORS

	Urban Modern	Urban Subsistence	Rural Subsistence
1) Type of employment	Modern nonmenial jobs in manufacturing and modern services	Domestic service, petty trades, handicrafts, repair menial labor services	Operators of very small-scale farms and landless laborers
2) Average productivity and income	High and growing	Low and stagnant	Low and stagnant
3) Barriers to employment	High: largely based on education and cultural background	Few if any	None

It can be seen that the number of years required for equalization of expected urban and expected rural earnings is higher for a given  $r$ , the lower is the value of  $\pi$ . And, for a given  $\pi$ , the higher the value of  $r$ , the further the break-even point moves into the future. It is interesting that the empirical literature spawned by Todaro's work gives little if any attention to the impact of  $r$ . Given the relatively high risks associated with LDC economic activity, the range of  $r = .05$  to  $r = .15$  would appear relevant. It becomes obvious that for many plausible situations, extremely long time horizons must be assumed if the Todaro model is to explain why a potential migrant with little education decides to take his chances in the urban modern labor market. If one must assume very long time horizons, in some cases greater than 50 years, an alternative explanation of migration may be in order.

#### IV

The following analysis is based on recognition that rural-urban migration is a dual phenomenon. Some migrants, those possessed of the requisite human capital, are bound for the  $U-M$  sector. Those who are less well endowed are intent on employment in the  $U-S$  sector. This recognition, in turn, requires some formal description of both the  $U-S$  and  $U-M$  sectors and of the rural sector from which the migrants are drawn. Table 3

presents the important features of the three relevant sectors. One sector of note in most developing countries is not shown in the table. Reference is to a modern rural sector composed of relatively large and relatively mechanized farms. Because the modern farms employ only a small proportion of the labor force and because the wage in that sector tends to be determined in the more populous rural subsistence sector, migration flows that originate or terminate in the modern rural sector have not been considered.<sup>16</sup> Although not formally incorporated into the model, the presence of a modern rural sector will be seen as important for the analysis.<sup>17</sup>

Contrasting with the advanced technology of the emerging  $U-M$  sector are the simple techniques of both subsistence sectors. Because of capital scarcity, the  $U-S$  sector and its rural counterpart feature low levels of productivity and incomes. And, as already discussed, the barriers to employment in the  $U-M$  sector are very high while there are few if any significant barriers to employment in either of the subsistence sectors.<sup>18</sup> This latter

<sup>16</sup>It is worth noting that the Todaro model apparently assumes that all of agricultural labor is found in a traditional or subsistence sector. He argues that institutional factors such as "traditional crop-sharing activities" dominate the rural sector (1969, p. 142, fn. 7).

<sup>17</sup>See fn. 41 below and the discussion to which it pertains.

<sup>18</sup>What is here called the  $U-S$  sector appears in other contexts in the literature and is labeled and defined variously. To some, the sector subsumes all employment not enumerated in official statistics. This often appears as a descriptive characteristic of what is called the informal sector (John Weeks, 1975). For some countries, however, and Mexico is an example, official statistics

It is apparent, however, that those many persons with no human capital would face zero probability of  $U-M$  employment.

fact means that the labor supply in the *U-S* sector is directly affected by conditions in the rural subsistence (*R-S*) sector.

As a basis for developing a formal model, it is necessary to recognize that a portion of *U-S* sector production is exported, so to speak, to other sectors while food and some manufactured goods are imported. For example, about 80 percent of all spending by low-income households in the Metroplex goes for food and clothing (Secretaria de Industria and Comercio, 1962, p. 111). Low-income urban households are not engaged in food production and it has been widely observed that clothing, footwear, and housewares from modern factories have increasingly been replacing the handicraft and artisan items traditionally produced within the *U-S* sector. These modern manufactured goods are imported from the *U-M* sector, and foodstuffs originate in the rural sectors, often passing through *U-M* intermediaries.

A capacity to import implies that the *U-S* sector also produces for export to other sectors. A formal conceptualization of the *U-S* sector must therefore recognize that some of that sector's employment produces goods and services to be consumed by households within the sector while the remainder produces exports for other sectors, largely the *U-M* sector. For the most part, those exports are in the form of menial labor services.<sup>19</sup> Even in

the case of handicrafts, so little capital is involved that value-added by the *U-S* sector is essentially due to the labor input. The *U-S* labor force therefore comprehends all urban persons employed at very low wages, whether found in the petty trades and handicrafts, in domestic service in *U-M* households, or as menials in otherwise modern factories or service establishments. A busboy at the Jockey Club is therefore categorized as a member of the *U-S* sector although his place of employment may appear conspicuously modern.<sup>20</sup> In this manner, the *U-S* labor force is made largely commensurate with urban poverty.<sup>21</sup>

## V

Because rural-urban migration is dual in nature, the migration rate is a composite of

---

*R-S* sector are largely produced in the *U-M* sector. There is, however, some derived demand for *U-S* labor, depending on the extent to which *U-M* firms employ menial labor. The rest of the world may import some of the output of handicraft industries and tourist use of hotel and restaurant facilities effects a derived demand for menial labor services. On the whole, however, the bulk of *U-S* sector exports are destined for the *U-M* sector and are largely in the form of labor services.

<sup>20</sup> For a detailed listing of occupations for Metroplex shanty town dwellers see Lomnitz (pp. 64-65). Some earnings data are illustrative. In the Metroplex, 65 percent of unskilled construction workers earned below the legal minimum wage in 1970, as did more than 90 percent of female service workers (Muñoz et al., p. 84).

<sup>21</sup> When it is recognized that certain employment categories in otherwise modern firms are of menial type, it is no longer reasonable to use the nature of the enterprise as the basis for delineating the modern and subsistence components of the labor force. A maid in a modern household is not considered a part of the modern labor force, nor is the force of grounds keepers at a modern manufacturing plant. The latter, using machetes to trim shrubs and shears to mow lawns, are essentially labor power substitutes for modern equipment just as maids are substitutes for modern appliances. This approach contrasts with convention. According to Sethuraman, "It is the enterprises and not the individuals in the urban economy that are classified into formal and informal sectors" (p. 71). Accordingly, a chambermaid in a modern hotel, earning a subsistence wage and residing in a shanty town, would be classified as a member of the formal or modern sector. Paulo Souza and Victor Tokman (1976, p. 359), however, apparently agree with us to the extent of including domestic servants as part of the informal sector.

---

attempt to enumerate all employment, even that of a marginal nature. For others, small-scale and/or family ownership are important distinguishing characteristics (S. V. Sethuraman, 1976). The informal sector is also taken to include those firms that operate either outside the law or in contravention of the law. This aspect appears in most definitions of the informal sector. Access to foreign exchange and relative labor intensity are also sometimes included in definitions of the sector. Barbara Harriss notes that "Almost all definitions [of the informal sector] contain arbitrary elements, run into tautology and have been found to be operationally controversial" (1978, p. 1077). Characteristics included by one definition are often ignored by others; none is sufficiently comprehensive. Moreover, none of these taxonomic systems provide the basis for a theoretical exploration of intersectoral relationships.

<sup>19</sup> The *R-S* sector is not a significant importer of labor services from the *U-S* sector, nor is the rest of the world. The manufactured goods, clothing, shoes, paper products, plastic wares, and the like, imported by the



migration into the two urban sectors. Thus,

$$(1) \quad (\dot{N}/N)(t) = (\dot{N}_m/N)(t) + (\dot{N}_{us}/N)(t)$$

recalling that  $N$  is the existing size of the urban labor force; and where  $\dot{N}$  represents net rural-urban migration. The subscripts  $m$  and  $us$  represent, respectively, the urban modern ( $U-M$ ) labor force and the urban subsistence ( $U-S$ ) labor force.

As did Todaro, we show migration into the  $U-M$  sector governed by the difference between the discounted streams of expected  $U-M$  and  $R-S$  real incomes, expressed as a percent of the discounted stream of expected rural income. Therefore,

$$(2) \quad (\dot{N}_m/N_m)(t) \\ = F[(V_m(t) - V_{rs}(t))/V_{rs}(t)], \quad F' > 0,$$

where  $V_m(t)$  is the discounted present value of the expected  $U-M$  real income stream over an individual's planning horizon; and  $V_{rs}(t)$  is the discounted present value of the expected real income stream in the  $R-S$  sector over the same planning horizon.

This is essentially the basic Todaro equation, but in the present context, it treats only those migrants bound for the  $U-M$  sector.<sup>22</sup>

For migrants who do not meet the human capital requirements of  $U-M$  work places, migration can be expressed as some function of the earnings differential between  $R-S$  and  $U-S$  sector employment.

$$(3) \quad (\dot{N}_{us}/N_{us})(t) \\ = F[(V_{us}(t) - V_{rs}(t))/V_{rs}(t)], \quad F' > 0,$$

where variables are defined in parallel fashion to those in equation (2). Equation (3) therefore explains the movement of all migrants who are not the subject of equation (2).

<sup>22</sup>This equation is very similar to equation (1) in Todaro's seminal article (1969). As in Todaro, it is assumed that all potential migrants have identical planning horizons; and, utilize constant and identical discount factors.

It will be noted that a common labor supply factor,  $V_{rs}$ , is found in both migration equations. A behavioral equation for that term can be formulated as follows:

$$(4) \quad V_{rs}(0) = \int_{t=0}^n Y_{rs}(t) e^{-rt} dt,$$

where  $Y_{rs}$  represents net income in the  $R-S$  sector (probably institutionally determined) in period  $t$ ,<sup>23</sup> and  $r$  is the appropriate consumption time-preference rate of discount.

By including  $V_{rs}$  as a labor supply factor for both streams of migrants, we are making the realistic assumption that the  $R-S$  wage tends to dominate the rural economy because few rural employment opportunities utilize human capital. We therefore consider that all potential rural emigrants use the  $R-S$  wage as their benchmark.

Focusing now on migrants bound for the  $U-M$  sector, we must consider an employment probability factor and also decide how to treat human capital. Because barriers to  $U-M$  employment are stated in terms of human capital, it is obviously a crucial factor. On the other hand, if the short run is of interest, investment in human capital is a sunk cost and therefore irrelevant to the migration decision. Where human capital would be useful is in explaining changes over time in migration rates. Our interest, however, like that of Todaro, will be the explanation of short-term migration decisions. Defining the short run as a period in which a potential migrant's stock of human capital is fixed and letting  $V_m(0)$  represent the present value of that person's perceived  $U-M$  earnings stream,

$$(5) \quad V_m(0) = \int_{t=0}^n P(t) Y_m(t) e^{-rt} dt - C(0),$$

where  $Y_m(t)$  represents net  $U-M$  real income in period  $t$ ,  $P(t)$  is the probability of having a  $U-M$  job in period  $t$ , and  $C$  is the one-time cost of the migratory move.

<sup>23</sup>See fn. 40 and equation (11) below.

The value  $V_m(0)$  therefore increases with increases in either  $P(t)$  or  $Y_m(t)$ , as in the Todaro case. If interest were in explaining migration over longer periods of time, human capital costs would be added to equation (5).<sup>24</sup>

The inclusion of a probability factor allows that at any point in time there may be an excess of  $U-M$  bound immigrants over  $U-M$  sector employment opportunities. Those educated migrants not initially obtaining  $U-M$  jobs, may either remain temporarily employed or find interim employment in the  $U-S$  sector. Based on considerable evidence it appears likely that many of the highly educated will accept temporary unemployment rather than take menial jobs.<sup>25</sup> To simplify the analysis we will assume that all  $U-M$  bound migrants who fail to find  $U-M$  employment, accept temporary unemployment.<sup>26</sup> Therefore, the probability factor  $P$  in equation (5) varies with the rate of open unemployment in the  $U-M$  sector.<sup>27</sup>

<sup>24</sup>In cases where the  $U-M$  sector is the focal point, the short-term migration decision presupposes a prior decision to invest in human capital. In its turn, that investment decision would have involved comparing expected future gains from migration with the opportunity cost of acquiring the requisite stock of human capital. From this view, the opportunity cost of education would be an important policy variable.

<sup>25</sup>Gunnar Myrdal develops this argument and concludes that "[the educated] are looking for nonmanual work and are not prepared to accept work that 'soils their hands'" (1968, pp. 1124 ff.).

<sup>26</sup>This assumption is necessary if we are to analyze the behavior of those migrants possessed of minimal human capital endowments. It has a strong basis in reality, however, being exemplified by the widespread phenomenon known as the educated unemployed. For those who have acquired relatively advanced schooling, employment in the  $U-S$  sector is considered demeaning. Further, those who could afford the opportunity cost of formal education are more likely to be able to afford temporary unemployment.

<sup>27</sup>Here  $P$  is related to  $\pi$  as shown in fn. 11, but, in this case,

$$\pi(t) = \gamma N_u(t) / (M(t) - N_u(t)),$$

where  $M$  is the total  $U-M$  labor force,  $N_u$  the employed  $U-M$  labor force, and  $\gamma$ , the rate of employment creation in the  $U-M$  sector (see Section III).

Because no human capital is required for employment in the  $U-S$  sector, the probability of employment at the existing  $U-S$  wage is unity for any given individual so that the present value of expected  $U-S$  sector earnings,  $V_{us}(0)$ , may be represented by

$$(6) \quad V_{us}(0) = \int_{t=0}^n Y_{us}(t) e^{-rt} dt - C(0),$$

where  $Y_{us}(t)$  represents net  $U-S$  real income in period  $(t)$ .

As in the Todaro case,  $U-M$  earnings,  $Y_m(t)$ , are determined by institutional factors.<sup>28</sup> However, because of the ease of entry, the  $U-S$  earnings,  $Y_{us}(t)$ , will tend to vary with the determinants of labor supply and labor demand. If we are to gain further understanding of the subsistence component of rural urban migration, we must pursue those determinants.

## VI

We follow Todaro's lead and consider labor supply to be exogenous.<sup>29</sup> Because of freedom of entry into the  $U-S$  sector and population pressure on rural land, labor supply for the  $U-S$  sector is considered to be perfectly elastic.<sup>30</sup> Analysis of labor demand requires specification of the determinants of output in the respective subsistence sectors. For the  $U-S$  sector, the production function is simple:

$$(7) \quad Q_{us} = F(N_{us}),$$

where  $Q_{us}$  and  $N_{us}$  denote output volume and size of the  $U-S$  labor force, respectively. Output is in the form of labor services and capital is assumed to be of scarcely any importance.

<sup>28</sup>Reference is usually to administered wages which are held to be significantly above the social opportunity cost of new hires.

<sup>29</sup>Population growth is a principle supply factor along with the labor force participation rate and the age structure of population. It would serve no useful purpose to include these in the model.

<sup>30</sup>For a discussion of the rural labor condition, see fn. 36 and the text to which it pertains.

In converting *U-S* sector output to income, two important identities emerge:

$$(8) \quad Y_{us} = P_{us}Q_{us} = \hat{W}_{us}N_{us}$$

$$(9) \quad Y_{us} = P_{us}(X + G),$$

where  $Y_{us}$ ,  $\hat{W}_{us}$ , and  $P_{us}$  are, respectively, total income, money wage, and price of labor services; and  $X$  and  $G$ , respectively, represent the volume of labor services sold as exports to the *U-M* sector and the volume of *U-S* output consumed within the *U-S* sector.<sup>31</sup>

The identity in equation (8) simply recognizes that because the output of the *U-S* sector is in the form of labor services, price and wage are different sides of the same coin. Equation (9) stipulates that a portion of the output of the *U-S* sector is exported, with the remainder consumed within the sector. This presents us with a useful perspective on *U-S* sector income. Indeed, if we assume that barter is inconsequential in the urban setting and that exports are the source of the circulating money supply in the *U-S* sector, the money value of exports largely determines the level of *U-S* sector income.<sup>32</sup> Whereas exports increase the income stream, imports of food and manufactured goods constitute leakages. Because there is little if any capital stock, we may assume there is no saving. The open economy multiplier for the *U-S* sector therefore depends solely upon the propensity to import. It then follows that the level of *U-S* sector income is based upon the money value of exports and the average propensity to import.<sup>33</sup> That income stream derived from

exports and expanded by a multiplier effect also finances consumption of some *U-S* sector value-added by *U-S* sector households.<sup>34</sup>

The most important determinants of demand for *U-S* exports are population and per capita income in the *U-M* sector, and the prices of capital goods that substitute for *U-S* labor. Where  $X$  represents the quantity of *U-S* exports demanded by the *U-M* sector, then

$$X = f(P_{us}, P_F, Y_m, Pop),$$

where  $P_{us}$  is the price of *U-S* labor service exports (*U-S* wage), and  $P_F$  represents the price of manufactured substitutes for *U-S* labor. Also,  $Y_m$  represents income per capita in the *U-M* sector and  $Pop$  stand for *U-M* population. If one were to understand the long-run impact of the *U-M* sector on *U-S* sector employment and rural-urban migration, the focus should be upon the nonprice variables.<sup>35</sup> At any point in time, however, the quantity demanded is a function of  $P_{us}$ , which, as we have seen, is the same as  $\hat{W}_{us}$ .

Given a demand schedule, any change in the *U-S* wage will affect the quantity of *U-S* exports and, depending on price elasticity, the total value of exports. This impact of price elasticity on the total value of exports, in turn, has implications for wage and employment. Indeed, *U-S* exports, *U-S* wage, and *U-S* employment are simultaneously determined. To establish this point it is necessary to include the *R-S* sector in the analysis

<sup>31</sup>A discussion of the export of labor services from the *U-S* sector was presented in Section IV.

<sup>32</sup>If we make the largely realistic assumption that the *U-S* sector is not privy to credit emanating from a fractional reserve system, then  $Y_{us} = \hat{X}V$ , where  $\hat{X}$  is the money value of exports, and  $V$ , the velocity of circulation.

<sup>33</sup>Indeed, if we assume balanced trade between the *U-S* sector and other sectors,  $Y_{us} = \hat{X}/APM$ , where  $\hat{X}$  is the money value of exports and  $APM$  represents the average propensity to import. (For changes in  $Y_{us}$ , the marginal propensity to import would be appropriate.)

Also note that  $\hat{X}V = Y_{us} = \hat{X}/APM$  (see fn. 32). It was noted earlier that 80 percent of expenditures by *U-S* households in the Metroplex went for imports. In that case, the export multiplier would be 1.25 and  $\hat{G}$  would be equal to .25 $\hat{X}$ . Other works have made reference to an export multiplier in a similar *LDC* context (Keith Hart, 1983), but the argument has not been rigorously developed.

<sup>34</sup>Because very little capital is involved, even that *U-S* sector value-added consumed by *U-S* sector households can be fairly said to consist very largely of labor input.

<sup>35</sup>An area of untapped research potential is the relationship of demand for labor services to the price of manufactured substitutes for those services.

because the absence of barriers to entry permits labor to move more or less freely between it and the *U-S* sector.

The output equation for the *R-S* sector can be represented as

$$(10) \quad Q_{rs} = f(N_{rs}, L_{rs}, K_{rs}),$$

where  $Q_{rs}$ ,  $N_{rs}$ ,  $L_{rs}$ , and  $K_{rs}$  denote, respectively, output volume, labor, land and capital. In the *R-S* sector, the stock of capital is minimal and possibilities for its expansion are severely circumscribed. Land is essentially fixed and, assuming that labor has a marginal product of zero over a wide range,<sup>36,37</sup>  $Q_{rs}$  would be at a maximum and remain so until land and/or capital bottlenecks were broken.<sup>38</sup>

We follow the standard practice of the literature on economic dualism and assume that the *R-S* real wage,  $W_{rs}$ , is institutionally determined, and is<sup>39,40</sup>

$$(11) \quad W_{rs} = Q_{rs}/N_{rs}.$$

<sup>36</sup>The often-used assumption of zero marginal product is usually based on the existence of a considerable pressure of population on the relevant land and an institutional arrangement that promotes application of family labor so long as returns are nonnegative. Severe population pressure exists in those regions of Mexico furnishing the bulk of rural to urban migrants, and in many other countries as well.

<sup>37</sup>Extreme assumptions such as zero marginal product are sometimes vital to the discovery of fundamental tendencies. Such discoveries, in turn, help to throw light on situations that do not necessarily conform to the extreme assumptions. For example, an empirical model based on the present theory could serve in situations where the marginal product of *R-S* labor is positive, albeit small relative to that of the *U-M* sector.

<sup>38</sup>This maximum is in reality a constraint. There could be shortfalls, of course, caused by natural phenomena. In that case the rural wage would fall, sparking additional rural-urban migration.

<sup>39</sup>This a, a real wage in the sense that it measures physical output and not in the sense that it measures constant purchasing power.

<sup>40</sup>The assumption that wage is equal to average product is reasonable for the collectives and small-scale private farms in Mexico where most labor comes from family members. The assumption can also apply to other Latin American or Asian countries with sharecropping systems by assuming that the return to land ownership is a fixed rent, with the remainder divided among members of the *R-S* labor force.

The money wage, which is used hereafter, is therefore

$$(12) \quad \hat{W}_{rs} = P_r(Q_{rs}/N_{rs}),$$

where  $P_r$  is the exogenously determined price of agricultural output which is assumed constant in the subsequent analysis,<sup>41</sup> and where labor supply is, in effect, perfectly inelastic.<sup>42</sup> With land a fixed input and the marginal product of labor equal to zero over a wide range,  $\hat{W}_{rs}$  will vary inversely with changes in  $N_{rs}$ . In this stylized situation, population growth places downward pressure on  $\hat{W}_{rs}$ . In the classical or Malthusian setting, a downward slide of the wage would ultimately act as a population check. However, in a dualistic setting such as that found in Mexico, there is a potential safety valve in the form of the *U-S* sector.<sup>43</sup>

With *U-S* labor and *R-S* labor considered substitutes, at least across a wide range, it follows that so long as the marginal product of labor in the *R-S* sector is at or near zero, the labor supply in the *U-S* sector will be perfectly elastic at a wage equal to that in the *R-S* sector. We may therefore combine equations (8) and (12) to arrive at the basic

<sup>41</sup>In Mexico, as in many other countries, there is a sizable modern rural sector. Agricultural prices, therefore, are not directly tied to particular changes in subsistence sector variables. It is therefore not desirable to include the intersectoral terms of trade as a variable in the model. A modern rural sector has appeared in previous dualism models (Keith Griffin, 1976).

<sup>42</sup>Population growth is the major supply factor and is taken to be exogenous. So long as institutional factors lead to employment of rural labor force members, the *R-S* labor supply for practical purposes will approach perfect inelasticity, while demand will feature unit elasticity in the relevant range. If, as assumed,  $Q_{rs}$  is at a maximum and  $P_r$  is exogenously determined, the value of *R-S* output is given at  $P_r Q_{rs}$ . In the assumed institutional setting this is also the fixed wage bill (from equation (2), we have  $P_r Q_{rs} = \hat{W}_{rs} N_{rs}$ ). Changes in  $N_{rs}$  are therefore precisely countered as  $\hat{W}_{rs}$  moves in the opposite direction.

<sup>43</sup>Illegal migration to the United States is also a safety valve for the Mexican population. However, a large proportion of the illegal migrants have significant stocks of human capital. For the relatively uneducated and the illiterate, the more likely safety valve is found in urban Mexico.

equilibrium condition shown in<sup>44</sup>

$$(13) \quad \hat{W}_{us} = Y_{us}/N_{us} = P_r(Q_{rs}/N_{rs}) = \hat{W}_{rs}.$$

By making certain assumptions about the costs of migration and discount rates, this equilibrium condition can be stated in terms of present value of expected earnings:  $V_{us} = V_{rs}$ .<sup>45</sup> Any reduction in  $\hat{W}_{rs}$  or increase in  $\hat{W}_{us}$  would therefore affect present value calculations (equation (3)) so as to increase the migratory flow.

Under the stipulated conditions, population pressure leading to labor force growth in either of the subsistence sectors will tend to reduce both the *R-S* and the *U-S* wage. There will be countervailing tendencies, however, the impact of which all depend on the relevant price elasticity of *U-S* exports.<sup>46</sup> To explore the importance of price elasticity, assume that the non-price-demand determinants for *U-S* exports are constant, and recall that  $P_{us}$  is also  $\hat{W}_{us}$ . As population pressure in the subsistence sectors pushes down the *U-S* wage, the quantity of exports demanded will increase, but, depending on the size of the elasticity coefficient, the income stream flowing into the *U-S* sector might swell or shrink. If *U-S* exports are price elastic, for example, the total value of labor services exported would increase as  $P_{us}$  falls, even if population and income in the *U-M*

sector did not grow.<sup>47</sup> In other words, the more price elastic the demand for *U-S* exports, the less will be the downward pressure on  $\hat{W}_{us}$  and the greater will be the expansionary influence on employment of a given labor supply pressure. The expansion of employment in the *U-S* sector is based partly upon the natural rate of population increase and partly upon rural-urban migration. On the other hand, if demand were inelastic, a falling *U-S* wage would reduce the total value of *U-S* labor service exports. In that case, supply pressure would cause a relatively large decrease in  $\hat{W}_{us} = \hat{W}_{rs}$ , and a relatively small increase in *U-S* sector employment.<sup>48</sup>

## VII

We have seen that wage and employment in the *U-S* sector are intimately tied to developments in other sectors. The *U-S* labor supply is greatly affected by conditions in the *R-S* sector and the demand for *U-S* labor largely emanates from the *U-M* sector. The exports associated with the demand for labor services result in a flow of payments into the *U-S* sector, the size of which is affected by the price elasticity of the labor exports.<sup>49</sup> The value of that flow, in turn, is

<sup>44</sup>Recall that  $Q_{rs}$  is at a maximum and that  $P_r$  is exogenously determined and assumed constant for the analysis. In reality the equilibrium might be achieved without equating money wages. We recognize that psychic return and/or cost of living might vary across sectors, but prefer to keep the analysis clear-cut by abstracting from such considerations.

<sup>45</sup>See equations (3), (4), and (6). The equilibrium  $\hat{W}_{us} = \hat{W}_{rs}$  would yield an equilibrium  $V_{us} = V_{rs}$  if the cost of migration were zero and time-preference discount rates were the same for rural and urban earnings. Given some cost of migration, when  $V_{us} = V_{rs}$ , then  $\hat{W}_{us} > \hat{W}_{rs}$ . Although it would complicate the exposition, the cost factor could be added to the analysis.

<sup>46</sup>In the classical surplus labor models, the rural subsistence wage determines the urban wage (W. Arthur Lewis, 1954). In the present model a common subsistence wage is determined by simultaneous changes in the *U-S* and *R-S* sectors.

<sup>47</sup>Having assumed *U-M* employment and income constant, the increased value of *U-S* exports would mean that *U-S* labor services have been substituted by the *U-M* sector for other goods and services. The *U-M* sector could increase its imports within the constraint of a constant level of income ( $Y_m = P_m Q_m$ ) so long as there were no savings in the *U-S* sector (the assumed condition). The *U-S* sector would therefore increase its imports of *U-M* sector products sufficiently to maintain balanced intersectoral trade, however, at a higher level.

<sup>48</sup>There have been virtually no attempts to estimate the price elasticity of the labor services of the type exported by the *U-S* sector. To the extent that the present analysis is useful, this topic should be added to the research agenda. In the absence of data, one is tempted to speculate that domestic service, for example, is price elastic, being in the nature of luxury.

<sup>49</sup>The extent to which growth in the *U-M* sector leads to *U-S* employment opportunities also depends on the income elasticity of *U-M* sector demand for *U-S* exports. There is a paucity of data on that subject. Dealing with combined demand of all urban sectors, one study suggests that the income elasticity coefficient varies from a low of less than 0.5 to well over 1.0 (Paul Mosley, 1978, p. 7). The data, however, did not treat

enhanced by a multiplier effect to create additional employment within the *U-S* sector.

Here then are the keys to understanding that portion of rural to urban migration that has resulted in burgeoning growth of urban subsistence sectors: 1) the growth of the subsistence population, which keeps downward pressure on the subsistence wage, and 2) the rapid growth of the *U-M* sector which has translated into the growth of demand for exports from the *U-S* sector. Where the Todaro model explains why those who possess human capital migrate, our model explains why masses of unschooled and relatively unskilled persons also join the trek to the city. They move when population pressure on fixed agricultural land reduces the rural subsistence wage significantly below that of the *U-S* sector, or when growth of demand for *U-S* exports pushes the wage in that sector significantly above that of the rural subsistence sector.<sup>50</sup> For many of those caught up in the urban trek, the focus is not on the modern sector with its relatively high wages, but rather on the subsistence sector with its relative ease of entry. The present analysis is therefore a useful complement to the Todaro theory of migration.

### VIII

Although the Todaro model is unidimensional while migration is a dual phenomenon, that model generally has been found consistent with the data. Consistency with the data, however, is not sufficient if the theory is inappropriate. It is the theory that informs policy. If the focus is growth of the *U-S* sector labor force, the Todaro model has yielded the correct prediction, but has failed to provide a satisfactory explanation.

---

domestic service or menial labor, categories crucial to this study.

<sup>50</sup>In reality, forces other than population also affect the rural wage. Various market and institutional forces may cause changes in prices for rural output. Recall, however, that we chose to hold rural prices constant in order to establish the basic relationships between the *R-S* and *U-S* sectors.

That theory, in turn, has lead analysts to the seemingly fatalistic view that attempts to solve the urban employment problem are likely to make the situation worse. This unintentional result is said to occur because the creation of new workplaces in the *U-M* sector enhances the probability of *U-M* employment, thereby promoting increased migration. Given the relatively large size of the rural labor force, the number of rural to urban migrants is likely to exceed the number of new modern sector workplaces. In that fashion, policies to promote urban economic modernization are said to lead to increased urban unemployment and underemployment (Todaro, 1979, p. 197). From that viewpoint, the market appears perverse. Decisions that are rational from the point of view of individuals appear disastrous for society. Those analysts are then forced to the conclusion that the solution to the urban employment problem must be sought in the rural areas. For example, it has been argued that "The present unnecessary economic incentives for rural-urban migration need to be minimized through creative and well-designed programs of integrated rural development" (Todaro, 1979, p. 198).<sup>51</sup> The perversion of the market is taken to be so extreme that some have been pushed to consider direct restrictions on migration such as those found in Tanzania and South Africa.<sup>52</sup>

<sup>51</sup>On the basis of the present analysis, it is appropriate to question whether investment directly in the rural sector would create ancillary employment of the nature and magnitude of that related to investment in the *U-M* sector. One is likely to find little demand for servants, plumbers, appliance repair shops, street vendors, news hawkers, boot blacks, or even small restaurants in agricultural villages. Finally, it is likely that the urban setting provides economies of scale in the delivery of education, health care, and family planning. There is the opportunity, therefore, for policymakers to take advantage of the rural to urban migration to reduce the population growth rate, on the one hand, and to improve the quality of the labor force, on the other. Rather than trying to stand against the tide, a reasonable policy emphasis might otherwise take advantage of the massive movement of humanity.

<sup>52</sup>While maintaining "grave reservations about the ethical issues involved," John Harris and Todaro (1970, p. 135) investigate the economic implications of direct restrictions on migration.

If one's viewpoint, however, includes the revised paradigm presented here, much of the recent rural-urban migration appears desirable from society's perspective as well as from that of individual migrants. The migration would be seen as socially undesirable only if one could show that a significant fraction of migrants produced in the urban setting less than would have been the case had they remained in rural environs. The facts appear to be otherwise, however. Because urban modern employment is likely for educated migrants, their movement contributes directly to economic growth.<sup>53</sup> In turn, the growth of the *U-M* sector fuels demand for *U-S* sector exports. The growth of the demand imparts a tendency for wage to increase in the *U-S* sector which, *pari passu*, sparks urban-bound migration into that sector. That migration itself improves the overall productivity of the economy and enhances the welfare of both those who migrate to the urban subsistence sector and those who remain in the rural subsistence sector.

<sup>53</sup>It can also be argued that those of the educated migrants who initially fail to find urban employment would not have been willing to work in the rural setting (see fn. 25). This might lead one to question the efficacy of having educated them, but that is another matter.

## REFERENCES

- Balán, Jorge, Browning, Harley and Jelin, Elizabeth, *Men in a Developing Society: Geographic and Social Mobility in Monterrey, Mexico*, Austin: University of Texas Press, 1973.
- Browning, Harley and Feindt, Waltraut, "Selectividad de Migrantes en un Metrópoli en un País en Desarrollo: Estudio de un Caso Mexicano," *Demografía y Economía*, No. 2, 1969, 3, 186-200.
- Gillis et al., Malcom, *Economics of Development*, New York: W. W. Norton, 1983.
- Griffin, Keith, *Land Concentration and Rural Poverty*, New York: Holmes and Meier, 1976.
- Harris, John R. and Todaro, Michael P., "Migration, Unemployment and Development: A Two-Sector Analysis," *American Economic Review*, March 1970, 60, 126-42.
- Harriss, Barbara, "Quasi-formal Employment Structures in the Unorganized Urban Economy, and the Reverse: Some Evidence from South India," *World Development*, September/October, 1978, 6, 1077-86.
- Hart, Keith, "Informal Income Opportunities and Urban Employment in Ghana," *Journal of Modern African Studies*, March 1983, 2, 61-89.
- Lewis, W. Arthur, "Economic Development with Unlimited Supplies of Labour," *Manchester School of Economics*, May 1954, 22, 131-91.
- Lomnitz, Larissa, *Networks and Marginality: Life in a Mexican Shanty Town*, New York: Academic Press, 1977.
- Lubell, Harold, "Urban Development and Employment: The Third World Metropolis," *International Labour Review*, November/December 1978, 117, 747-56.
- McGee, T. G., *The Urbanized Process in the Third World*, London: G. Bell and Sons, 1971.
- Mosley, Paul, "Implicit Models and Policy Recommendations: Policy Towards the Informal Sector in Kenya," *IDS Bulletin*, February 1978, 3-10.
- Muñoz et al., Humberto, *Migración y Desigualdad Social en la Ciudad de México*, Mexico City: El Colegio de México, 1977.
- Myrdal, Gunnar, *Asian Drama*, Vol. III, New York: Pantheon, 1968.
- Peattie, Lisa R., *The View from the Barrio*, Ann Arbor: University of Michigan Press, 1968.
- Santos, Milton, *The Shared Space: The Two Circuits of the Urban Economy in Underdeveloped Countries*, London: Methuen, 1979.
- Sethuraman, S. V., "The Urban Informal Sector: Concept, Measurement and Policy," *International Labour Review*, July/August 1976, 114, 69-81.
- Souza, Paulo and Tokman, Victor, "The Informal Sector in Latin America," *International Labour Review*, November/December 1976, 114, 355-65.
- Sundrum, R. M., *Development Economics*, New York: Wiley & Sons, 1983.
- Todaro, Michael P., "A Model of Labor Mi-

- gration and Urban Unemployment in Less Developed Countries," *American Economic Review*, March 1969, 59, 138-48.
- \_\_\_\_\_, *International Migration in Developing Countries: A Review of Theory, Evidence Methodology, and Research Priorities*, Geneva: International Labour Office, 1976.
- \_\_\_\_\_, *Economic Development of the Third World*, New York: Longman, 1979.
- Weeks, John, "Policies for Expanding Employment in the Informal Urban Sector of Developing Economics," *International Labour Review*, January 1975, 111, 1-13.
- Secretaría de Industria y Comercio, Departamento de Muestreo, *Las 16 Ciudades Principales de la República Mexicana: Ingresos y Egresos Familiares*, Mexico City, 1962.
- \_\_\_\_\_, Dirección General de Estadística, *IX Censo General de Población*, Mexico City, 1972.



# Accounting Rates of Return

By GERALD L. SALAMON\*

Many economic studies provide empirical evidence regarding cross-sectional differences in firm profitability. In most of these studies, firm profitability is measured by an accounting rate of return (net income dividend by the book value of assets, hereafter, *ARR*) rather than an economic rate of profit.<sup>1</sup> Franklin Fisher and John McGowan define the firm's economic rate of profit (*IRR*) as that interest rate which "equates the present value of its net revenue stream to its initial outlay" (1983, p. 82). Fisher and McGowan examine the analytic relation between a firm's *ARR* and its *IRR* in a series of examples and conclude that the *ARR* is such a bad surrogate for the *IRR* that the results of *ARR*-based empirical studies are likely to be "totally misleading" (p. 91).

William Long and David Ravenscraft (1984) have criticized the theoretical work of Fisher and McGowan because the nature of the analytic relationship between the *ARR* and the *IRR* in the context of highly simplified hypothetical "examples" may not be indicative of the nature of the relationship in empirical settings. Their criticism adopts the utilitarian view that the *ARR* has to be treated as a suitable surrogate for the *IRR* as long as no preferable alternative measure of profitability is available for empirical work.

This view has merit as long as the measurement error which is contained in the *ARR* is random rather than systematic. Unfortunately, the nature and extent of the measurement error which is contained in the *ARR* in a particular empirical setting can only be determined unequivocally if the *IRR* is unequivocally known. In such a case, of course, there would be no need to rely on the *ARR* at all. Due to this unhappy circle, empirical research on firm profitability has continued to rely on the *ARR* despite the fact that many persons reasonably believe that the results of such research may be totally misleading (Fisher and McGowan; G. C. Harcourt, 1965), while others reasonably believe that the results are reliable enough to form a basis for policy decisions (Long and Ravenscraft).

Recent work by Y. Ijiri (1978; 1979; 1980) and by myself (1982) has shown that conditional *IRR* estimates can be obtained from data in firms' financial statements. While these *IRR* estimates are conditional, they do abstract from certain extraneous factors that influence *ARR*s and that differ across firms. Consequently the conditional *IRR* estimates are free from *some* of the sources of measurement error which are known to contaminate the *ARR*. Thus, the conditional *IRR* estimates can be used to provide evidence on whether these sources of measurement error have or have not affected the outcome of prior studies in which profitability was measured by the *ARR*. Such evidence can help to *objectively* resolve the conflict between the opposing views on the reliability of *ARR*-based profitability research.

This paper uses conditional *IRR* estimates to examine the properties of the measurement error in the *ARR* in a study of the relationship between firm profitability and firm size. The remainder of the paper is organized as follows. In Section I, the theoretical work of Fisher and McGowan is used

\*Professor of Accounting, University of Iowa, Iowa City, IA 52242. I gratefully acknowledge the comments of William Albrecht, Dan Dhaliwal, Gary Fethke, Franklin Fisher, Scott Linn, William R. Kinney, Jr., and the participants of a workshop at the University of Florida on earlier versions of this paper.

<sup>1</sup>Some authors (for example, William Long and David Ravenscraft, 1984) have used the sales margin ratio (earnings/sales) as a measure of profitability. However, if one views profitability as return per unit of sacrifice, the sales margin ratio is not a profitability measure since it ignores the sacrifice (or investment) required to generate a dollar of sales. Consequently, this paper does not attempt to shed any insight into prior studies that have relied on the sales margin ratio as a measure of profitability.

as the basis for analyzing the nature of the measurement error in the *ARR* in general and in the context of studies of the relationship between firm profitability and firm size. An outline and an analysis of the procedures for estimating conditional *IRRs* is presented in Section II. In Section III, empirical evidence is presented that documents that the *ARR* contains systematic rather than random measurement error in firm size studies. Finally, in Section IV, a brief summary of the paper is provided and some specific suggestions for future research are made.

### I. Measurement Error and Firm Size

G. Whittington (1979, p. 204) has argued that the *ARR* would be a suitable surrogate for the *IRR* in empirical research if the *ARR* is correlated with the *IRR*, and if the variance of the *ARR* that is unexplained by the *IRR* is uncorrelated with the explanatory variable used in the analysis. If such were the case, then the use of the *ARR* in empirical research would lead to weak but unbiased tests of economic hypotheses. Research that has relied on conditional *IRR* estimates suggests that the *ARR* and the *IRR* are positively associated in most empirical settings (for example, my 1982 article; myself and M. Moriarty, 1984; my 1984 paper with S. C. Linn and W. P. Albrecht). Consequently, the first of Whittington's criteria for satisfactory surrogation appears to be generally met. Unfortunately, it is still unclear whether the second of his conditions is or is not met in general. The essence of the controversy between Fisher and McGowan and their critics (for example, Long and Ravenscraft) appears to center on this single issue. Fisher and McGowan appear to hold the position (based upon their theoretical analysis) that the error in the *ARR* is likely to be systematic in enough settings that the use of the *ARR*, in general, produces results that cannot be relied upon. Long and Ravenscraft, on the other hand, seem to believe that the measurement error in the *ARR* is likely to be random in most settings so that tests of economic hypotheses which rely on the *ARR* are unbiased and thus reliable. At this stage of our knowledge, it is difficult to evaluate the merits of these opposing positions be-

cause of the lack of direct evidence. Furthermore, neither of these two extreme positions is likely to hold in general. In particular, it is likely that the measurement error in the *ARR* will be random in some settings and systematic in other settings. If this is the case, then it would be quite useful to establish guidelines that would allow a preliminary classification of past *ARR*-based research into one set in which the major concern would be the power of the statistical tests and into another set in which the major concern would be the bias of the statistical tests. This paper represents an initial effort to base such guidelines on the outcome of empirical tests of the theory developed by Fisher and McGowan.

Fisher and McGowan demonstrate that the level of a firm's *ARR* is systematically affected by its choice of a depreciation method. Thus, their theory suggests that the nature of the measurement error in the *ARR* in a particular sample of firms will depend upon the distribution of accounting methods across the firms in that sample. In particular, the measurement error in the *ARR* is likely to be systematically related to a potential explanatory variable whenever the choice of accounting methods is systematically related to that variable.

This paper examines the nature of the measurement error in the *ARR* in the context of studies of the relationship between firm profitability and firm size. This area was chosen for examination for two primary reasons. First, the relationship between firm size and firm profitability is a time-honored topic on which well-known scholars have differing opinions. Marshall Hall and Leonard Weiss (1967, p. 329) suggest that William Baumol (1959) believed in the existence of a positive association between firm size and firm profitability while George Stigler (1950) and Joe Bain (1956) expressed either skepticism or uncertainty about such a positive association. Hall and Weiss present empirical results that are consistent with a positive association between firm size and profitability, but profitability in their study was measured by the *ARR*.

The second, and perhaps more important, reason for examining the nature of the *ARR*-based measurement error in the context

of a firm-size study is that prior research has found that there are systematic differences in the accounting methods adopted by firms of different size. For example, R. Watts and G. Zimmerman (1978), D. Dhaliwal, myself, and E. D. Smith (1982), and R. Hagerman and M. Zmijewski (1979) all found that large firms tended to adopt accounting methods that resulted in smaller earnings levels than the earnings levels produced by the accounting methods adopted by small firms.<sup>2</sup> This means, for example, that large firms tend to elect accelerated methods of depreciation more frequently than small firms. The nature of these differences in the depreciation methods of large and small firms is important given Fisher and McGowan's demonstration of how a firm's *ARR* is affected by its choice of a depreciation method. In particular, Fisher and McGowan (Table 2, p. 86) demonstrated that, if the growth rate in investment is less than the *IRR* of firm projects, then the steady-state *ARR* of the firm is higher if it uses an accelerated depreciation method than if it uses the straight-line depreciation method.<sup>3</sup> Thus, the nature of the systematic differences in the depreciation methods adopted by large and small firms is hypothesized to induce systematic rather than random measurement error in the *ARR* in the context of studies of the relationship between firm profitability and firm size.

It is important to notice that the direction of this effect is such that the positive association between firm size and firm profitability that was observed by Hall and Weiss may have been induced by the systematic error contained in their measure of profitability the *ARR*. Consequently, an examination of the properties of the measurement error in

the *ARR* in size-profitability studies has the potential to provide evidence that directly addresses the controversy between Fisher-McGowan and their critics (as represented by Long-Ravenscraft). Clearly, the theory developed by Fisher and McGowan along with the findings of systematic differences in the accounting methods adopted by large and small firms predicts that the *ARR* will contain systematic measurement error in studies of the relationship between firm profitability and firm size. If such systematic error is not documented in this setting then it would appear that the *ARR* is a reasonable general surrogate for the *IRR* in empirical research as was suggested by Long-Ravenscraft. On the other hand, if the *ARR* is found to possess measurement error properties consistent with the predictions of the theory developed by Fisher-McGowan, then their theory cannot be legitimately criticized because of the simplicity of its underlying assumptions. Furthermore, if systematic measurement error is found to be present in the *ARR* in this setting, then the results of *ARR*-based research should be viewed with skepticism whenever there is an association between the explanatory variable of concern and the accounting methods used by the firm.

## II. Conditional Estimates of Firm *IRR*

Ijiri (1978; 1980) has advocated the preparation of financial reports on the cash basis rather than on the accrual basis. Ijiri's emphasis on cash-basis reports has stimulated interest in a variable called the cash recovery rate (*CRR*). The *CRR* is the ratio of the firm's cash recoveries in a period to the gross historical cost of investments outstanding during the period. Under the condition of constant investment growth, my article (1982) has shown that a firm's *CRR* converges to the constant given in equation (1) when the cash flow profile of the firm's projects can be represented by a single parameter.<sup>4</sup>

<sup>2</sup>The studies of accounting method choice and firm size often report significance levels between 5 and 10 percent. See R. Holthausen and R. Leftwich (1983) for a thorough review of the accounting method choice literature.

<sup>3</sup>The case of *IRR* being less than the rate of investment growth is the most common case in practice for the large mature firms which are examined in this paper. See my 1973 article for a demonstration that shareholders are continually increasing their investment in the firm (i.e., they never receive a dividend) whenever the growth rate exceeds the *IRR*.

<sup>4</sup>This equation is similar to equation (4) of my article (1982). Equation (1) above, however, emphasizes the nominal *IRR* rather than the real *IRR*.

$$(1) \quad CRR = g / [(1 + g)^n - 1] \\ \cdot \left[ \frac{(1 + g)^n - b^n}{(1 + g - b)} \right] \cdot \left[ \frac{(1 + r)^n (1 + r - b)}{(1 + r)^n - b^n} \right],$$

where  $g$  = constant growth rate in gross investment,  $n$  = useful life of firm's "representative" composite project,  $r$  = internal rate of return of firm's representative project, and  $b$  = project cash flow pattern parameter. If  $Y_0, Y_1, \dots, Y_n$  are the after-tax cash flows of the firm's representative projects with  $Y_0 < 0$  and  $Y_1, \dots, Y_n > 0$ , then  $b$  is such that  $Y_i = b^{i-1} Y_1$  for  $i = 1, \dots, n$ . Note that if  $b < 1$  ( $>$ ) project cash flows decay (grow) exponentially.

Certain assumptions about the firm and its environment underlie the derivation of equation (1). In particular, each firm is assumed to be a collection of projects that have the same useful life, cash flow profile, and *IRR*. The collection of projects is assembled via a gross investment function that grows at a constant rate per period. It is true, of course, that contrary to these assumptions, firms do not have a constant rate of investment growth and they do invest in a variety of projects with different lives, cash flow profiles, and *IRRs*. However, there are arguments and evidence that suggest that the assumptions made in order to derive equation (1) are not so unrealistic as to prevent it from being used to obtain meaningful empirical *IRR* estimates. For example, Ijiri (1979, p. 261) has observed that a firm can be thought of as a composite project with a given life and cash flow profile as long as the underlying mix of its individual projects is reasonably stable over time. The stable project mix assumption would seem to be a reasonable one for large mature firms since for such firms the impact of new ventures would seem to be relatively small. Additionally, Ijiri (1979) has provided indirect evidence on the issue of cyclical vs. constant rates of investment growth in mature firms. Significant cyclical patterns in investment growth would lead to cyclical patterns in *CRRs*.<sup>5</sup> Ijiri (1978, Table

2, p. 341, and fn. 12, p. 342) presents evidence that the *CRRs* for a sample of mature firms are reasonably time stable. Thus, it appears that the assumptions which underlie equation (1) are not systematically and importantly violated if its use is restricted to samples of mature firms.

Equation (1) allows the empirical estimation of the *IRR* of an on-going firm given estimates of *CRR*,  $g$ ,  $n$ , and  $b$ . The *CRR*,  $g$ , and  $n$  can be reasonably estimated from data in the firm's external financial reports. The cash flow pattern parameter,  $b$ , poses a more difficult problem. In principle, this parameter could be estimated from a regression of cash recoveries on a distributed lag of past investment (Fisher-McGowan, p. 91, fn. 21). However, very little empirical work has been conducted on the cash flow profiles of firm projects because of econometric difficulties associated with distributed lag regressions and because the empirical research on how *IRR* estimates are affected by cash flow profiles is at such an early state of development.

Prior work has sidestepped the problem of estimating cash flow profiles by assuming that all firms have projects with the same cash flow pattern parameter which is assigned a numerical value (see my 1982 article). The fact that the cash flow pattern parameter is assigned a value (and is not estimated) means that the estimates of *IRR* that result from solving equation (1) for  $r$  have to be viewed as conditional rather than as unconditional estimates. While the decision to assign a value to the cash flow pattern parameter does avoid the estimation problem, it does raise questions about how to choose the value which is assigned to the parameter. In a preliminary examination of this issue (1982), I found that the rankings of firms by their conditional *IRRs* was relatively insensitive to the value assigned to the cash flow profile parameter in equation (1). It is not clear, however, whether the insensitivity of *IRR* rankings to different levels of the cash flow pattern parameter will mean that the reported measurement error properties of the *ARR* in a particular setting will also be insensitive to whether the firms' *IRRs* are estimated with different values of the cash flow pattern parameter. Consequently,

<sup>5</sup> See J. L. Livingstone and myself (1970) for a model which leads to cyclical yet convergent *ARRs*.

in this paper, four different conditional *IRRs* were calculated for each firm—each different *IRR* was based upon a different value of the cash flow pattern parameter. The values assigned to the cash flow parameter (*b*) were .8, 1.0, 1.1, and a random assignment from a uniform distribution which was defined on the interval (.8, 1.1). This practice allows this paper to present some preliminary evidence on whether the reported measurement error properties of the *ARR* are sensitive to the cash flow profiles which are assumed when equation (1) is used to convert firms' *CRRs* into estimates of their *IRRs*.<sup>6</sup>

I argued (1982, p. 301) that typical after-tax cash flow patterns would be declining because of the widespread use of accelerated depreciation methods for tax purposes and because of the use of the payback period as a criterion for choosing projects. Consequently, one conditional estimate of the *IRR* of each firm was based on the assumption that project cash flows were declining. For this case, the parameter *b* was assigned the value of .8. A value of *b* less than .8 was not examined because of the fact that most of the firms in the sample had reasonably long project lives. Extremely declining cash flow patterns and long project lives are inconsistent with the choice of an optimal depreciable life for tax purposes. Assigning *b* a value of 1 corresponds to level cash flows over the projects' lives and also results in equation (1) being considerably simplified. In particular, if *b* is equal to 1 then the firm's *CRR* converges to the constant given in

$$(2) \quad CRR = r / [1 - (1 + r)^{-n}].$$

Ijiri (1978; 1980) and myself and Moriarty relied on equation (2) to obtain conditional *IRR* estimates for samples of mature firms. It is noted in this case of level cash flows that the growth rate in the firm's gross investment has no influence on the level of the firm's *CRR*. The cash flow pattern parameter was also assigned the value 1.1 in order to allow

for the case of an increasing cash flow pattern over the life of the projects. So far the assignment of values to the cash flow pattern parameter was done under the assumption that all firms had projects which had the same cash flow pattern. Consequently, while this procedure allowed an examination of how sensitive the results were to different common cash flow patterns, it did not allow a determination of whether the results would be affected if different firms had different cash flow patterns. In order to examine this latter issue, a value was assigned to the cash flow pattern parameter of each firm by taking a random drawing from a uniform probability distribution which had end points of .8 and 1.1. Thus, in this paper four different conditional *IRRs* were estimated for each firm in the sample with each estimate corresponding to a different assumption about the cash flow pattern of the firms' projects. The results of this paper were quite similar for all four conditional *IRR* estimates and consequently only the results based upon the conditional *IRR* estimates which assumed common and level cash flows (labeled *IRR*(1)) and which made a random assignment to the value of the cash flow pattern parameter (labeled *IRR*(ran)) are presented.

In closing this section, it is noted that the *ARR* contains measurement error due to the fact that its level is influenced by the length of project life, the growth rate in gross investment, and the accounting methods adopted by the firm (for example, see Fisher-McGowan)—factors which will be different for different firms in a particular sample. The conditional *IRR* estimate, because they rely on cash flow data, explicit estimates of project lives and investment growth rates, are not influenced by these factors extraneous to profitability assessment. Consequently, the conditional *IRR* estimates are profitability estimates which are free from three specific sources of measurement error which are known to be present in the *ARR*.

### III. The Evidence

#### A. Data Description

In order to examine the nature of the measurement error in the *ARR* in the con-

<sup>6</sup>I am grateful to an anonymous reviewer of a previous version of this paper for the comment that led to the incorporation herein of the "sensitivity" issue.

text of studies of the relationship between firm profitability and firm size, the data to calculate the *ARR*, four conditional *IRRs*, and a measure of firm size was collected. The firms which were the subject of analysis were the manufacturing firms on a COMPUSTAT tape which possessed all the financial statement data required to calculate the values of these variables for each of the five years 1974–78. There were 197 firms which met the data requirements.<sup>7</sup>

The size of each firm (hereafter *TA*) was measured as the average of its Total Assets (net) over the period 1974–78. The *ARR* for each firm for each year was calculated as Net Income before Interest Expense, Extraordinary Items, and Discontinued Operations divided by the average Total Assets (net) for the year. The simple average of a firm's five annual *ARRs* was taken as the *ARR* of the firm for the period 1974–78.

Equation (1) was used to estimate four conditional *IRRs* for each firm in the sample. The estimates were obtained by following the procedures described below. First, a *CRR* for each firm for each year was calculated as cash recoveries divided by the gross investment for the year. Cash recoveries were calculated as the sum of Total Funds from Operations, Interest Expense, Sale of Investments, Sale of Property, Plant, and Equipment, and the Decrease in Current Assets (if it occurred). This calculation of cash recoveries followed that of Ijiri (1978). The calculation of cash recoveries does not directly depend upon the level of or the change in the firm's liabilities or equities because of the desire to disentangle the results of the firm's investment activities from the results of its financing activities. The calculation also does not make a distinction between the short- and long-term investments of the firm. Increases in all assets are treated as investments (rather than as reductions in cash recoveries) and decreases in all assets are associated with cash recoveries (i.e., recoveries associated with a prior investment of funds).

Gross investment for each year was calculated as the average of beginning and ending Total Assets (gross) (See Ijiri, 1978, pp. 345–47; 1980, pp. 55–56). The simple average of a firm's five *CRRs* was defined to be the firm's *CRR* for the period 1974–78. Second, an estimate of the useful life of the projects and of the investment growth rate of each firm was needed in order to convert (via equation (1)) each firm's *CRR* into conditional estimates of its *IRR*. An estimate of the useful life of firm projects was obtained for each firm for each year by dividing the average of its Gross Plant for the year by its Depreciation Expense for the year. The average of these annual estimates was used as the estimate of the life of each firm's projects. An estimate of each firm's growth rate in gross investment over the period 1974–78 was calculated as  $1/5 \log (\text{Gross investment, 1978} / \text{Gross investment, 1974})$ . These steps gave values for *CRR*, *n*, and *g* for each firm in the sample. Finally, these estimated values along with the previously discussed values for the parameter *b* (i.e., *b* = .8, 1.0, 1.1, or a random assignment from (.8, 1.1)) were substituted into equation (1) and produced four conditional *IRR* estimates for each firm in the sample. Table 1 presents some summary statistics for the variables that are used in the remaining analysis of this paper—*ARR*, *IRR*(1), *IRR*(*ran*), and *TA*.

### B. Empirical results

The first step in the data analysis was to determine the properties of the measurement error contained in the *ARR* and to determine if the nature of the estimated properties was sensitive to the assumed level of the firm's cash flow parameter. This determination began with a cross-sectional regression of *ARR* on each conditional *IRR* estimate. Information regarding the cross-sectional regression of *ARR* on *IRR*(1) is presented in Panel A of Table 2 while information about the regression of *ARR* on *IRR*(*ran*) is presented in Panel B of Table 2. In both cases the *ARR* and the conditional *IRR* estimate are significantly and positively related. It is important to note, however, that there is considerable variation in the *ARR* that is not

<sup>7</sup>A computer listing of these firms and their industry affiliation is available from the author upon request.

TABLE 1—SAMPLE SUMMARY STATISTICS: PROFITABILITY MEASURES AND AVERAGE SIZE 1974–78, 197 FIRMS

	ARR	IRR(1) <sup>a</sup>	IRR(ran) <sup>b</sup>	TA
A: Univariate Statistics				
Mean	8.96	8.70	9.26	1.84
Standard Deviation	4.88	5.71	6.69	4.03
Minimum	-10.00	-13.8	-14.4	.03
Maximum	25.00	28.3	31.3	38.29
B: Correlation Matrix				
ARR	1.00			
IRR(1)	.64	1.00		
IRR(ran)	.51	.93	1.00	
TA	.16	.09	.08	1.00

Note: ARR, IRR(1), and IRR(ran) are shown in percent; TA is hundred-millions of dollars.

<sup>a</sup>b = 1.0.  
<sup>b</sup>b = ran.

TABLE 2—THE ASSOCIATION OF ARR AND CONDITIONAL IRR, GIVEN LEVEL PROJECT CASH FLOWS FOR ALL FIRMS (IRR(1)) AND RANDOMLY ASSIGNED FIRM-SPECIFIC PROJECT CASH FLOWS (IRR(RAN)) 197 FIRMS

Variable	Estimated Coefficient	Standard Error	t-Score	Significance Level
A: $ARR = \beta_1 IRR(1) + \beta_0 + ME(1)$				
IRR(1)	.545	.047	11.67	.001
Constant	.042	.005	—	—
Correlation (ARR, IRR(1)) = .64				
B: $ARR = \beta_1 IRR(ran) + \beta_0 + ME(ran)$				
IRR(ran)	.374	.045	8.3	.001
Constant	.055	.005	—	—
Correlation (ARR, IRR(ran)) = .51				

explained by the estimated conditional IRRs. In particular, approximately 60 percent (74 percent) of the cross-sectional variation in the ARR is explained by factors other than the profitability of firm projects as measured by IRR(1) (IRR(ran)).

The regressions of ARR on each of the conditional IRRs were used to obtain an estimate of the measurement error in the ARR of each firm. The ARR-based measurement error was calculated as follows:

(3)  $ME_j(.) = ARR_j - [\beta_1 IRR_j(.) - \beta_0]$ ,

where  $ME_j(.)$  = an estimate of the measurement error in the ARR of firm  $j$  ( $j = 1, 2, \dots, 197$ ) given a conditional  $IRR_j(.)$  estimate of firm  $j$  with  $IRR_j(.) = [IRR_j(1),$

$IRR_j(ran)]$ ;  $\beta_1, \beta_0$  = the estimated slope and intercept coefficients from the cross-sectional regression of ARR on  $IRR(.)$ ; and  $ME_j(.)$  represents an estimate of that part of the ARR of firm  $j$  which is influenced by factors other than the conditional  $IRR_j(.)$  of firm  $j$ 's projects. The level of the association between  $ME_j(.)$  and TA will determine whether the ARR does or does not contain systematic measurement error in the context of studies of the relationship between firm profitability and firm size.

Table 3 displays the results of the regressions of  $ME_j(.)$  on firm size (TA). Panel A displays the results given the conditional IRR that assumed a common and level cash flow pattern for the projects of all firms. Panel B displays the results given the conditional IRR

TABLE 3—THE ASSOCIATION OF *ARR* MEASUREMENT ERROR AND FIRM SIZE (*TA*)<sup>a</sup>

Variable	Estimated Coefficient	Standard Error	t-Score	Significance Level
A: $ME(1) = \beta_1 TA + \beta_0$				
<i>TA</i>	.0012	.00066	1.88	.05 <sup>b</sup>
Constant	-.2284	.2921	—	—
Correlation ( $ME(1), TA$ ) = .134				
B: $ME(ran) = \beta_1 TA + \beta_0$				
<i>TA</i>	.0014	.0007	1.94	.05 <sup>b</sup>
Constant	-.2624	.3257	—	—
Correlation ( $ME(1), TA$ ) = .138				

<sup>a</sup> $ME(1)$  assumes level project cash flows for all firms and  $ME(ran)$  assumes randomly assigned firm-specific project cash flows.

<sup>b</sup>One-tailed level.

TABLE 4—FIRM PROFITABILITY AND FIRM SIZE (*TA*)<sup>a</sup>

Variable	Estimated Coefficient	Standard Error	t-Score	Significance Level <sup>b</sup>
A: $ARR = \beta_1 TA + \beta_0$				
<i>TA</i>	.0019	.0009	2.25	.025
Constant	8.61	.3780	—	—
Correlation ( $ARR, TA$ ) = .159				
B: Level Project Cash Flows $IRR(1) = \beta_1 TA + \beta_0$				
<i>TA</i>	.0012	.0010	1.23	.15
Constant	8.47	.4468	—	—
Correlation ( $IRR(1), TA$ ) = .088				
C: Random Project Cash Flows $IRR(ran) = \beta_1 TA + \beta_0$				
<i>TA</i>	.0013	.0012	1.12	.15
Constant	8.97	.5236	—	—
Correlation ( $IRR(ran), TA$ ) = .080				

<sup>a</sup>Profitability measured by *ARR*, *IRR(1)*, and *IRR(ran)*.

<sup>b</sup>One-tailed level.

estimate that was based upon a random assignment of a different value to the cash flow pattern parameter of each firm. In both cases, there is a significant ( $\alpha = .05$ ; one-tailed test) positive association between the estimated measurement error in the *ARR* and the size of the firm.

The results depicted in Panels A and B of Table 3 are important for three reasons. First, given that the positive association between the *ARR*-based measurement error and firm size was robust with respect to the four different project cash flow profiles examined, it would be difficult to attribute the systematic nature of the *ARR*-based measurement error to the fact that the cash flow profiles in this paper were assumed rather

than estimated. Second, the observed positive association between  $ME_j(.)$  and *TA* is consistent with the differences in the accounting methods used by large and small firms and the predicted impact of such differences on the reported *ARR*. Consequently, this result can be interpreted as a partial validation of Fisher-McGowan's theoretical model. Third, the fact that the measurement error in the *ARR* in firm-size studies is systematic rather than random means that the tests of association between *ARR* and *TA* are biased rather than weak tests of the association between firm profitability and firm size. The next step in the data analysis was to determine if the strength of the association between  $ME_j(.)$  and *TA* was strong



enough to induce an importantly higher association between firm profitability and firm size when profitability is measured by the *ARR* than when profitability is measured by a conditional *IRR*.

Table 4 presents the results of regressing three measures of firm profitability on firm size. Panels B and C of Table 4 reveal that profitability and size are not significantly related when profitability is measured by *IRR*(1) or *IRR*(ran), respectively. On the other hand, Panel A of Table 4 reveals that profitability and size are significantly ( $\alpha = .025$ , one-tailed test) positively related when profitability is measured by the *ARR*. The results of Tables 3 and 4, taken together, are consistent with the view that it is the nature and the strength of the measurement error in the *ARR*, rather than the correctness of the underlying economic arguments, which accounts for the observed positive association between firm size and firm profitability in studies which have measured profitability by the *ARR*.

#### IV. Summary and Suggestions for Future Research

This paper has relied upon conditional *IRR* estimates to demonstrate that the *ARR* contains systematic—not random—measurement error in the context of studies of the relationship between firm profitability and firm size. The results of this paper suggest two different avenues for future research. The first avenue concerns the determination of the nature of the measurement error in the *ARR* in studies where explanatory variables other than firm size are the subject of analysis. Systematic measurement error is likely to be found in some of these studies because of the dependence of the *ARR* on the accounting methods adopted by firms and the documented systematic association between some economic variables and these accounting methods. For example, Dhaliwal, myself, and Smith found systematic differences in the depreciation methods adopted by management-controlled and owner-controlled firms. Consequently, studies of the relationship between *ARR* and the control status of firms (for example, David

Kamerschen, 1968, and R. Joseph Monsen, J. S. Chiu, and D. E. Cooley, 1968) are predicted to be ones in which the *ARR* will contain systematic rather than random measurement error. On the other hand, the *ARR* is likely to contain random measurement error in some settings because there is only a weak or no association between the explanatory variables of interest and the accounting methods used by firms. For example, there is mixed evidence about whether firms in industries of different sales concentration have systematically different accounting methods (for example, see Hagerman and L. Senbet, 1976, vs. Hagerman and Zmijewski) and there is no evidence to suggest that firms of different advertising intensity have systematically different accounting methods. Consequently, the *ARR* is predicted to have the properties which make it a reasonable *IRR* surrogate in studies of industrial sales concentration and advertising intensity. Preliminary evidence consistent with these predictions is provided in my paper with Moriarty and my paper with Linn and Albrecht. Economists, by following this first avenue of research, will begin to develop a body of evidence which will allow past *ARR*-based studies that are reliable to be conceptually separated from those that are not.

The second avenue for future research arises because the *IRR* estimates relied upon in this paper are conditional on assumed knowledge of the cash flow profiles of firm projects. A broad-based study involving regressions of cash recoveries on a distributed lag of past investment as suggested by Zvi Griliches (see Fisher-McGowan, p. 91, fn. 21) would provide empirical evidence on the cash flow profiles of actual firms. This evidence would provide guidance to analytic work on the relationship between a firm's *CRR* and *IRR* so that future empirical research would be able to rely on more refined *IRR* estimates than those used in the current study.

#### REFERENCES

- Bain, Joe S., *Barriers to New Competition*,  
Cambridge: Harvard University Press,

- 1956.
- Baumol, William J., *Business Behavior, Value, and Growth*, New York: Macmillan, 1959.
- Dhaliwal, D., Salamon, G., and Smith, E. D., "The Effect of Owner Versus Management Control on the Choice of Accounting Methods," *Journal of Accounting and Economics*, July 1982, 4, 41-53.
- Fisher, Franklin M., and McGowan, John J., "On the Misuse of Accounting Rates of Return to Infer Monopoly Profits," *American Economic Review*, March 1983, 73, 82-97.
- Hagerman, R. and Senbet, L., "A Test of Accounting Bias and Market Structure," *Journal of Business*, October 1976, 49, 509-13.
- \_\_\_\_\_, and Zmijewski, M., "Some Economic Determinants of Accounting Policy Choice," *Journal of Accounting and Economics*, August 1979, 1, 141-61.
- Hall, Marshall and Weiss, Leonard W., "Firm Size and Profitability," *Review of Economics and Statistics*, August 1967, 49, 319-31.
- Harcourt, G. C., "The Accountant in a Golden Age," *Oxford Economic Papers*, March 1965, 17, 66-80.
- Holthausen, R. and Leftwich, R., "The Economic Consequences of Accounting Choice: Implications of Costly Contracting and Monitoring," *Journal of Accounting and Economics*, August 1983, 5, 77-117.
- Ijiri, Y., "Cash-Flow Accounting and Its Structure," *Journal of Accounting, Auditing, and Finance*, Summer 1978, 1, 341-48.
- \_\_\_\_\_, "Convergence of Cash Recovery Rate," in his and A. Whinston, eds., *Quantitative Planning and Control*, New York: Academic Press, 1979, 259-67.
- \_\_\_\_\_, "Recovery Rate and Cash Flow Accounting," *Financial Executive*, March 1980, 48, 54-60.
- Kamerschen, David R., "Ownership and Control and Profit Rates," *American Economic Review*, June 1968, 58, 432-47.
- Livingstone, J. L., and Salamon, G. L., "Relationship Between the Accounting and the Internal Rate of Return Measures: A Synthesis and an Analysis," *Journal of Accounting Research*, Autumn 1970, 8, 199-216.
- Long, William F. and Ravenscraft, David J., "The Misuse of Accounting Rates of Return: Comment," *American Economic Review*, June 1984, 74, 494-500.
- Monsen, R. Joseph, Chiu, J. S. and Cooley, D. E., "The Effect of the Separation of Ownership and Control on the Performance of the Large Firm," *Quarterly Journal of Economics*, August 1968, 82, 435-51.
- Salamon, Gerald L., "Models of the Relationship Between the Accounting and Internal Rates of Return: An Examination of the Methodology," *Journal of Accounting Research*, Autumn 1973, 11, 296-303.
- \_\_\_\_\_, "Cash Recovery Rates and Measures of Firm Profitability," *Accounting Review*, April 1982, 57, 292-302.
- \_\_\_\_\_, Linn, S. C., and Albrecht, W. P., "The Effect of Measurement Error in the Accounting Rate of Return on the Estimated Relationship Between Firm Profitability and Industry Concentration," unpublished paper, University of Iowa, 1984.
- \_\_\_\_\_, and Moriarty, M., "Alternative Profitability Measures and Tests of Hypotheses: An Application to the Advertising-Profitability Issue," unpublished working paper, University of Iowa, 1984.
- Solomon, E., "Return on Investment: The Relation of Book Yield to True Yield," in R. K. Jaedicke et al., eds., *Research in Accounting Measurement*, Menasha: American Accounting Association, 1966, 232-44.
- Stauffer, Thomas R., "The Measurement of Corporate Rates of Return: A Generalized Formulation," *Bell Journal of Economics*, Autumn 1971, 2, 434-69.
- Stigler, George, "Monopoly and Oligopoly by Merger," *American Economic Review Proceedings*, May 1950, 40, 23-34.
- Watts, R. and Zimmerman, G., "Towards a Positive Theory of the Determination of Accounting Standards," *Accounting Review*, January 1978, 53, 112-34.
- Whittington, G., "On the Use of the Accounting Rate of Return in Empirical Research," *Accounting and Business Research*, Summer 1979, 9, 201-08.

# Urban Land Prices under Uncertainty

By SHERIDAN TITMAN\*

Land prices in west Los Angeles are among the highest in the United States. Yet, we can observe a number of vacant lots and grossly underutilized land in this area. A good example of this is a parking lot, owned by the University of California-Los Angeles, in an area of Westwood where land has been known to sell for more than \$100 per square foot. The university could probably raise a considerable amount of money by selling two-thirds of the parking lot and constructing a parking structure on the remaining land to satisfy the demand for parking. Although this may be one of the best examples of underutilized land in west Los Angeles, it is by no means the only example. There are many underutilized and vacant urban lots throughout Los Angeles and the rest of the world, held by private investors who presumably wish to maximize their wealth.

The fact that investors choose to keep valuable land vacant or underutilized for prolonged periods of time suggests that the land is more valuable as a potential site for development in the future than it is as an actual site for constructing any particular building at the present time. Hence, in order to understand why certain urban lots remain vacant, we must determine how the land is valued under the two alternatives. Valuing the land as a site for constructing a particular building at the current time is fairly straightforward. It is simply the market value of the building (including the land) minus the lot preparation and construction costs (this is referred to in the real estate literature as residual value). However, valuing the vacant land as a potential building site is not as straightforward since the type of building

that will eventually be built on the land, as well as the future real estate prices, are uncertain.

The model developed in this paper provides a valuation equation for pricing vacant lots of this type. Although the model is very simple, it provides strong intuition about the conditions under which it is rational to postpone building until a future date. Furthermore, the pricing model can be adapted to provide realistic estimates of urban land values in much more complex settings.

The notion that it is often optimal to delay irreversible investment decisions has previously been considered in the environmental economics and capital investments literature.<sup>1</sup> The basic intuition in these papers is that it may be advantageous to wait for additional information before deciding upon the exact specifications of the investment project. While the authors demonstrate that it is often valuable to delay investment, and maintain the option to choose a better investment project in the future, they do not explicitly show how this option affects the value of other related assets in their models.

This paper adapts the methods first used by Fisher Black and Myron Scholes (1973) and Robert Merton (1973), to value options and other derivative securities, to determine explicit values for vacant urban land. The valuation model is particularly close in its approach to the binomial option pricing models of John Cox, Stephen Ross, and Mark Rubinstein (1979), and Richard Rendleman and Brit Bartter (1979). The intuition being that a vacant lot can be viewed as an option to purchase one of a number of different possible buildings at exercise prices that are equal to their respective construction costs.

\*Graduate School of Management, University of California, Los Angeles, CA 90024. I thank Fred Case, Nai-Fu Chen, Margaret Fry, Mark Grinblatt, Frank Mittelbach, and Brett Trueman for helpful comments.

<sup>1</sup>See, for example, John Krutilla (1967), Alex Cukierman (1980), Douglas Greenley, Richard Walsh, and Robert Young, (1981), and Ben Bernanke (1983).

This approach provides a valuation formula that is a function of observable variables and is independent of the investor's preferences.

This valuation technique should be contrasted to the standard textbook approach to valuing vacant land under uncertainty.<sup>2</sup> Richard Ratcliff (1972), for instance, suggests that appraisers determine the most probable future use of the land, appraise the property as of that future time and that use, and then discount this future value to the present. This method ignores the fact that the type of building that will be constructed in the future is generally unknown, and will be determined by real estate prices at that time. The analysis in this paper demonstrates that the amount of uncertainty about the type of building that will be optimal in the future is an important determinant of the value of vacant land. If there is a lot of uncertainty about future real estate prices, then the option to select the type of building in the future is very valuable. This makes the vacant land relatively more valuable and makes the decision to develop the land at the current time relatively less attractive. However, if there is very little uncertainty about future real estate prices, the option to select the appropriate type of building in the future is relatively less valuable. In this case, the decision to develop the land at the current time is relatively more attractive.

My analysis provides more than just a novel method for valuing land under uncertainty. It enables us to address issues, previously unexplored, that pertain to the effect of uncertainty on real estate markets. My results relating to how uncertainty about future real estate prices affect current real estate activities has important policy implications. For example, the analysis suggests that government action intended to stimulate con-

struction activities may actually lead to a decrease in such activities if the extent and duration of the activity is uncertain. The analysis also has policy implications regarding the imposition of height restrictions on buildings. It is shown that the initiation of height restrictions, perhaps for the purpose of limiting growth in an area, may lead to an increase in building activity in the area because of the consequent decrease in uncertainty regarding the optimal height of the buildings, and thus has the immediate affect of increasing the number of building units in an area.

The paper is organized as follows: Section I examines the type of building, characterized by its size, that will be built at a given date if the land is to be developed at that time. Section II presents a simple two-date, two states of nature, model for determining the value of the vacant land for the case where the future price of building units, and hence the size of the building that is to be constructed, is uncertain. A simple numerical example that illustrates this valuation technique is presented in Section III. Section IV presents a comparative static analysis of this valuation model which includes, among other things, an analysis of the effect of uncertainty on vacant land value. Section V examines a model where the current price and rental rate on building units as well as land values are endogenous and Section VI provides a numerical example which illustrates how the valuation technique can be applied to value land with many possible building dates and many possible states of nature corresponding to each date.

### I. The Optimal Building Size

Buildings, in this model, are characterized by their size, or number of units,  $q$ . The cost of constructing a building on a given piece of land,  $C$ , is an increasing and convex function of the number of units, that is,  $dC/dq > 0$  and  $d^2C/dq^2 > 0$ . The rationale for the second assumption is that as the number of floors in a building increases, labor costs per floor increase and the foundation of the building must be stronger. It is also assumed that subsequent to completing a building of

<sup>2</sup>I am unaware of any extant land pricing models that consider uncertainty. However, Donald Shoup (1970), Chapman Findlay and Hugh Howson (1975), and James Markusen and David Scheffman (1978) have examined some of the issues analyzed here within certainty models. Also, René Stulz (1982) suggested that the model he developed for pricing options to purchase one of two risky assets could be applied to price land in some specific cases.

a certain size, it is prohibitively expensive to add additional building units.

Given these assumptions, the building size that maximizes the wealth of a landowner who wishes to construct a building at the present time will satisfy the following maximization problem:

$$(1) \quad \underset{(q)}{\text{Max}} \Pi(p_0) = p_0 q - C(q),$$

where  $p_0$  is the current market price per building unit.

Differentiating (1) with respect to  $q$ , it follows that the solution to this maximization problem is to choose a building size which satisfies the condition,

$$(2) \quad dC/dq = p_0.$$

The building size that satisfies this equality will be denoted  $q^*$ . Given this optimal decision, it follows directly that the value of the land for building at the present time,  $\Pi(p_0)$ , is an increasing and convex function of  $p_0$ . It should be noted that the convexity results because the landowner can change  $q^*$  in response to changes in  $p_0$ .

I will later demonstrate, within a more specialized model, that because of this convexity property, greater uncertainty about the future unit price of buildings increases the current value of vacant land. The basic intuition behind this result can be seen by comparing the expected value of the land for building at date 1, over uncertain realizations of  $\tilde{p}_1$  with the value of the land given a known date 1 price of  $\hat{p}_1 = E(\tilde{p}_1)$ . It follows from Jensen's inequality that

$$(3) \quad E(\Pi(\tilde{p}_1)) > \Pi(E(\tilde{p}_1)).$$

Hence, uncertainty increases the expected future value of the vacant land. This implies that uncertainty causes current vacant land values to increase at least for the case where investors are risk neutral.<sup>3</sup>

<sup>3</sup>For the special cases where  $\tilde{p}_1$  is normally distributed, or where  $C(q)$  is quadratic, the expected future value of land is monotonically increasing in the variance of  $\tilde{p}_1$ .

## II. Valuing Urban Land under Uncertainty

Here I present a simple model for valuing land under uncertainty. Although the model makes no assumptions about investor preferences, other simplifying assumptions are made. The model consists of only two dates, so if the landowner chooses not to build at the present date (date 0), he or she will develop the land at date 1 if  $\pi(p_1) > 0$ . Holding vacant land is assumed to generate no revenues or costs. Uncertainty, in this model, enters in a very simplistic manner. First, the only source of uncertainty pertains to the market price of building units. Per unit construction costs are known and constant. Furthermore,  $\tilde{p}_1$ , the date 1 price of building units takes on one of only two possible values,  $p_h$  and  $p_l$ , where  $p_h > p_l$ . Given that building units can take on only two possible prices on the second date and building costs are constant, it follows that the land at date 1 can take on only two possible values,  $\pi(p_h)$  and  $\pi(p_l)$ . It should be noted that these simplifying assumptions are relaxed considerably in Section VI. It is also assumed that a risk-free asset exists with a return of  $R_f$ . The per unit rental rate,  $R_t$ , is initially assumed to be exogenous; however, this variable is determined endogenously in the model presented in Section V. Finally, it is assumed that markets are perfect in that there are no taxes, no transaction costs, and no short-selling restrictions.<sup>4</sup>

The vacant land can be considered what the finance literature refers to as a contingent or derivative security. Its date 1 value is completely determined by (or derived from)

<sup>4</sup>The assumption of frictionless markets, generally assumed in models of security prices, is considered by some to be less realistic when applied to real estate markets. However, it should be noted that securities represent indirect claims on factories and equipment that are probably much less liquid than real estate. Yet we can price these assets as if they were perfectly liquid because the securities are traded on (almost) frictionless markets. Similarly, a large fraction of real estate is held by publicly traded firms. If the real estate investments of these firms are chosen in a manner consistent with value maximization, then real estate prices will be determined in equilibrium as if markets were really frictionless.

an exogenously priced asset, the date 1 price of building units. In the finance literature, options and other contingent securities are valued by forming a hedge portfolio, consisting of the risk-free asset and the exogenously priced primitive asset, that is perfectly correlated with the contingent security. In the absence of riskless arbitrage, the contingent security must have the same value as this hedge portfolio.

The vacant land can be similarly valued in this model. Since there exist three investments (land, building units, and the risk-free asset) that take on at most two possible values, the returns of the vacant land can be exactly duplicated by a linear combination of the returns of the building units and the risk-free asset. Hence, in the absence of riskless arbitrage, the price of the vacant land can be determined as a function of these investments.

An easy way to solve this pricing problem is to first determine the state prices, (i.e., the cost at date 0 of receiving one dollar in one of the two date 1 states of nature and zero dollars in the other), and then sum the products of these state prices and the land values in the two states of nature. These state prices,  $s_h$  and  $s_l$ , must satisfy the following two equations that express the date 0 price of building units and the price of a discount bond as functions of their date 1 cash flows:

$$(4) \quad p_0 = s_h p_h + s_l p_l + R_f (s_h + s_l)$$

$$(5) \quad 1/(1 + R_f) = s_l + s_h.$$

Solving these equations yields the following state prices for high and low price states of nature, respectively:

$$(6) \quad s_h = \frac{p_0 - (p_l + R_f/1 + R_f)}{p_h - p_l}$$

and

$$(7) \quad s_l = \frac{(p_h + R_f/1 + R_f) - p_0}{p_h - p_l}.$$

Given these state prices, it follows that if no opportunities for riskless arbitrage exists,

the price of vacant land at date 0 must be

$$(8) \quad V = \Pi(p_h) s_h + \Pi(p_l) s_l.$$

If the value of the vacant land, as specified in equation (8), exceeds the profit from building at the present date,  $\Pi(p_0)$ , the wealth-maximizing landowner will choose to have the land remain vacant. Otherwise, he or she will build at date 0 the size building that satisfies equation (2).

### III. A Simple Numerical Example

Consider the example where an investor owns a lot that is suitable for either six or nine condominium units. The per unit construction costs of the building with six and nine units is \$80,000 and \$90,000, respectively. The current market price of the units is \$100,000. The per year rental rate is \$8,000 per unit (net of expenses) and the risk-free rate of interest for the year is 12 percent. If market conditions are favorable next year, the condominiums will sell for \$120,000; if conditions are unfavorable, they will sell for only \$90,000.

Since the marginal cost, per unit, of building nine rather than six units is \$110,000, the investor will build a six-unit building and realize a profit of \$120,000 if he builds at the current time. However, if he chooses to wait one year to build, he will construct a six-unit building if market conditions are unfavorable and realize a total profit of \$60,000, and will build a nine-unit building and realize a total profit of \$270,000 if favorable market conditions prevail. Substituting these numbers into equation (8) yields a current value for this land, if it is to remain vacant until next year, of \$141,071. Since this is greater than the profit that would be realized by building immediately, it is better to keep the land vacant.

If the land sells for less than this amount, investors can earn arbitrage profits by purchasing the land, and hedging the risk by short-selling the condominium units. For example, if the land sold for \$120,000, investors could realize a risk-free gain with no initial investment by purchasing the land, short-selling seven condominium units, and in-

vesting the net proceeds from the transactions in the risk-free asset. The seven condominium units completely hedges the risk from owning the vacant land since the difference between the value of the units in the good and bad states of nature, \$210,000, exactly offsets the difference in land values in the two states. Hence, the above investment yields a risk-free gain of \$23,600. Since such gains cannot exist in equilibrium, investors will bid up the price of the land to its equilibrium value of \$141,071.

#### IV. Comparative Statics

The above numerical example illustrates the effects of the current price of the building units, the interest rate, and the rental rate on the current value of vacant land. Recall that in order to hedge the risk from owning the vacant land, individual building units were sold, with the proceeds invested in the risk-free asset. If the price of the building units increases, the proceeds from the short sale increase, so the vacant land becomes more valuable. Similarly, if the interest rate increases, the income from the risk-free asset increases so the vacant land becomes less valuable. Conversely, if the rental rate increases, the cost of maintaining the short position increases, so the value of the vacant land decreases.

These comparative static results can be shown formally by differentiating equation (8) under the assumption that  $p_h$  and  $p_l$  are fixed:

$$(9a) \quad \partial V / \partial p_0 = \frac{\Pi(p_h) - \Pi(p_l)}{p_h - p_l} > 0,$$

$$(9b) \quad \partial V / \partial R_f = \frac{\Pi(p_h)(p_l + R_f) - \Pi(p_l)(p_h + R_f)}{(p_h - p_l)(1 + R_f)^2} > 0,$$

$$(9c) \quad \partial V / \partial R_t = \frac{\Pi(p_l) - \Pi(p_h)}{(p_h - p_l)(1 + R_f)} < 0.$$

The preceding analysis implicitly assumes that the current price and rental rate on

building units are unaffected by changes in the risk-free rate. Alternatively, we can examine the case where  $R_t$  is constrained to equal  $R_f p_0$ . A change in the risk-free rate accompanied by a proportional change in the rental rate can then be analyzed by substituting  $R_f p_0$  for  $R_t$  in equation (8) to yield

$$(8') \quad V = \Pi(p_h) \left( \frac{p_0 - p_l}{(p_h - p_l)(1 + R_f)} \right) + \Pi(p_l) \left( \frac{p_h - p_0}{(p_h - p_l)(1 + R_f)} \right).$$

It is clear from the above equation that the value of the vacant land decreases if an increase in interest rates is accompanied by a corresponding increase in rental rates.

The valuation technique presented in Section IV above also enables us to analyze the effect of increased uncertainty on land values. This is done by considering the effect of increasing the spread between  $p_h$  and  $p_l$  in such a way that state prices remain constant, and are consistent with both current rental rates and the prices of building units remaining constant. Hence, the effect of uncertainty on land values established here is applicable to cross-sectional comparisons holding current building prices constant.

One can easily verify that if  $p_h$  increases by  $x$  dollars and  $p_l$  decreases by  $xs_h/s_l$  dollars, the state prices remain unchanged. Also, the value  $p_h s_h + p_l s_l = p_0 - R_t / (1 + R_f)$  remains unchanged. This is consistent with, but does not require,  $p_0$  and  $R_t$  to remain unchanged. However, the value of vacant land,

$$(10) \quad V = \Pi(p_h + x)s_h + \Pi(p_l - (xs_h/s_l))s_l,$$

is an increasing function of  $x$ . This can be seen by differentiating  $V$ , in equation (10), with respect to  $x$ :

$$dV/dx = \Pi'(p_h + x)s_h + \Pi'(p_l - (xs_h/s_l))s_h.$$

It follows from the convexity of  $\Pi(p)$  that

$$dV/dx > 0 \quad \text{since}$$

$$\Pi'(p_h + x) > \Pi'(p_l - (xs_h/s_l)).$$

This result indicates that if the amount of uncertainty increases, the value of the vacant land increases, decreasing the relative attractiveness of constructing a building at the current time. Developing the land at the current time becomes less attractive because the increased uncertainty about future prices makes the size of the building that will be optimal at the future date more uncertain, which in turn makes it more likely that the optimal building size at the current time will be suboptimal in the future. If the building size ( $q^*$ ) that will be constructed in the future is known, perhaps because of height restrictions, then the amount of uncertainty about future prices will not enter the decision as to whether to build now or to build in the future. The decision will instead be determined by a comparison between the rental rate and the return from investing the construction expenses in the risk-free asset. This can be seen by comparing the value of the land for constructing a building with  $q^*$  units at the current time period:

$$(11) \quad \Pi = p_0 q^* - C(q^*),$$

with its value as a building site for next period:

$$(12) \quad V = s_h [p_h q^* - C(q^*)] + s_l [p_l q^* - C(q^*)].$$

Substituting equation (7) into (12) yields

$$(13) \quad V = p_0 q^* - R_l q^* (s_h + s_l) - C(q^*) (s_h + s_l),$$

which suggests that the building should be constructed at the present date if and only if,

$$(C(q^*) + R_l q^*) / (1 + R_f) > C(q^*),$$

which simplifies to

$$(14) \quad R_l q^* > R_f C(q^*).$$

Since condition (14) is less restrictive than the condition  $\Pi(p_0) > V$  (for the case where there are no building restrictions), a particular piece of land may be developed at the present date (if height restrictions are imposed), in circumstances under which it would not be developed otherwise. Hence, the imposition of height restrictions can conceivably have the immediate effect of increasing the number of building units in a particular area.

The effects of changes in future building prices, which do lead to changes in current building prices, can also be examined within this model. An increase in  $p_h$ , holding  $p_l$ ,  $s_l$ ,  $s_h$ , and  $R_l$  constant, will increase  $p_0$  by the amount  $s_h$  (see equation (4)), which in turn will increase the profit from developing the land at the current date by the amount

$$(15) \quad d\Pi/dp_h = \Pi'(p_0) s_h.$$

From equation (8), this increase in  $p_h$  leads to an increase in the value of the vacant land of

$$(16) \quad dV/dp_h = \Pi'(p_h) s_h.$$

If  $p_h$  exceeds  $p_0$ ,  $\Pi'(p_h)$  will exceed  $\Pi'(p_0)$  since  $\Pi(\cdot)$  is convex. In this case, an increase in building prices in the good state of nature increases the current value of the vacant land relative to its value if developed. Hence, it becomes less attractive to build at the current date. In the less likely case where the price of building units in the favorable state of nature is lower than the current price, an increase in  $p_h$  makes it more attractive to build at the current date.

Similarly, a decrease in  $p_l$ , holding the other variables constant, decreases current building unit prices by  $s_l$ , which in turn leads to a decrease in the profit from developing the land at the current date by the amount

$$(17) \quad d\Pi/dp_l = \Pi'(p_0) s_l.$$



This decrease in  $p_l$  leads to a corresponding decrease in the vacant land value of

$$(18) \quad dV/dp_l = \Pi'(p_l)s_l.$$

It follows, from the above equations, that a decrease in  $p_l$  will lead to a decrease in the profit from developing the land at the current date that is greater (less) than the corresponding decrease in the value of the vacant land if  $p_0$  exceeds (is less than)  $p_l$ . The above analysis suggests that any increase in the  $p_h - p_l$  spread makes it relatively more valuable to delay developing the land as long as  $p_h > p_0 > p_l$ . This conforms to the basic intuition that increased uncertainty increases the value of having open alternatives. However, this intuition does not necessarily hold when either  $p_0 > p_h$ , or  $p_0 < p_l$ .

#### V. A Simple Examination

Here I present a simple examination of the effect of increased uncertainty on equilibrium prices and building activity. Up to this point, I have not addressed issues relating to the effect of uncertainty on the current prices and rental rates of building units. In order to do this, I must add structure to the model. The following analysis examines a simple economy that consists of  $N$  identical lots that are initially vacant. If, in equilibrium, all the lots are developed at date 0, then there will exist no vacant lots to value. Conversely, if none of the lots are developed, no building units will exist. Hence, it makes sense to restrict the analysis to equilibria in which some, but not all, of the lots are developed at date 0. This suggests that, in equilibrium, the date 0 value of a vacant lot must equal the profit from developing it at that time:

$$(19) \quad V_0 = \Pi(p_0).$$

The demand for building units at date 0 is expressed as a decreasing function of their rental rate:

$$(20) \quad Q = nq^* = f(R_l),$$

where  $Q$ , the number of building units demanded, is equal to the product of  $n$ , the number of lots that are developed in the current period, and  $q^*$ , the number of building units constructed per lot. The function  $f(R_l)$  is assumed to be continuous and differentiable with  $df/dR_l$  less than zero.

Equations (1), (2), (4), (8), (19), and (20), along with the exogenous  $p_l$ ,  $p_h$ , and  $R_f$ , define a well-specified equilibrium.<sup>5</sup> The effect of uncertainty on this equilibrium can be explored in the manner developed in the previous section; by increasing  $p_h$  by  $x$  and decreasing  $p_l$  by  $xs_h/s_l$  so that  $p_0 - (R_l/(1 + R_f))$ ,  $s_h$  and  $s_l$  remain unchanged.

As was shown previously, an increase in uncertainty of this type leads to an increase in  $V$ . This implies that  $\Pi(p_0)$  must increase, which in turn implies that both  $p_0$  and  $q^*$  must increase. Since  $p_0 - (R_l/(1 + R_f))$  remains constant with changes in  $x$ ,  $R_l$  must also increase. From equation (20) we see that  $Q$  decreases with increases in  $R_l$ . Since  $q^*$  increases and  $Q$  decreases, it must be the case that  $n$  decreases. In other words, if uncertainty is increased in a manner that keeps the state prices constant, prices of both land and building units as well as rental rates will increase, a larger portion of the land will remain vacant, but taller buildings will be constructed.

#### VI. Extensions and Practical Applications

Because of tractability considerations, the valuation model developed in Section II was kept simple. The model consisted of only two dates, with only two possible states of nature at the second date, and construction costs were assumed to be fixed. While these assumptions allow us to easily analyze the effects of uncertainty on land prices, they can be relaxed if our only interest is in developing a practical technique for valuing urban land.

<sup>5</sup>Note that the above equations are all continuous and that the variables are all finite and nonnegative. Hence, the existence of this equilibrium follows directly from Brouwer's fixed-point theorem.

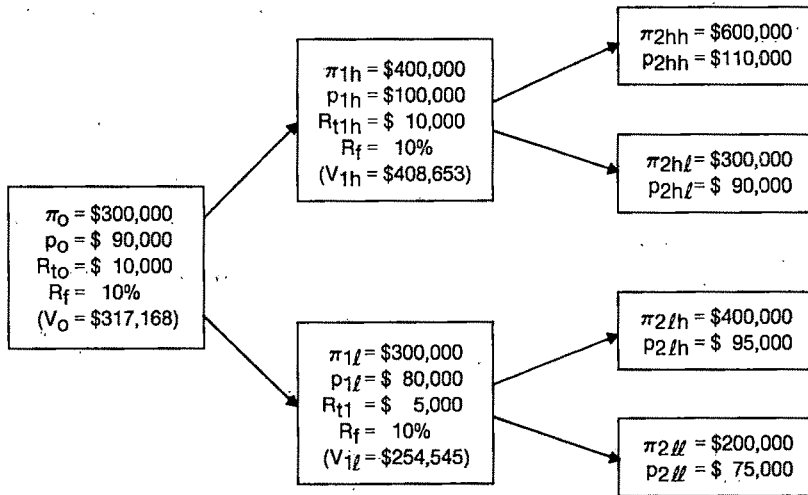


FIGURE 1

The assumption that construction costs are certain can easily be relaxed. The profit from constructing the optimal size building in each date and state of nature can be calculated as long as the construction costs and the per unit price of buildings is specified for each date and state. Substituting these profit levels into equation (8) yields the value of the vacant land.

The pricing model can also be generalized to allow for more than two dates. This can be done by specifying that for each date  $t$  state of nature, two possible date  $t+1$  states of nature can occur. The date 0 land value can then be solved by backwards induction. For each state of nature at the second to last date, the vacant land value is given by equation (8). The larger of this value and the profit from developing the land in each state of nature at this date can then be substituted for  $\Pi$  into equation (8) to calculate the values of vacant land at the third to last date for the different states of nature. By continuing this process, we not only obtain the current value of the vacant land, but also determine at which future dates and states of nature the land is developed. Note also that by making the time periods between dates arbitrarily small and the number of dates arbitrarily large, we can have an arbitrarily large number of states of nature for each future time period. Hence, the assumption of

only two date  $t+1$  future states of nature for each date  $t$  state is not really restrictive.

The following numerical example illustrates this valuation method. It assumes three dates. The profit from developing the land, the per unit building price, and the rental rate is given for each date and state of nature in Figure 1. The value of the vacant land in the two date 1 states of nature are calculated in the manner specified in Section II. Since the value of the vacant land in the favorable state of nature (\$408,635) exceeds the profits from developing the land in this state of nature, the land will remain vacant. However, the value of the land is only \$254,545 in the unfavorable date 1 state of nature. Since this value is less than the profit from developing the land at that date, the land will be developed if the unfavorable state of nature occurs. Substituting the larger of the value of the vacant land and the profit from developing the land in each state of nature for  $\Pi(p)$  in equation (8) yields the date 0 value of the vacant land. Since this value (\$317,168) exceeds the profit from developing the land at date 0, the land will remain vacant at this date.

## VII. Conclusion

The model developed in this paper provides a valuation equation for pricing vacant

lots in urban areas. The analysis demonstrates that the range of possible building sizes provides a valuable option to the owner of vacant land that becomes more valuable as uncertainty about future prices increases. An implication of this relationship between uncertainty and vacant land values is that increased uncertainty leads to a decrease in building activity in the current period.

The relationship between building activity and uncertainty may have important macro implications. An article by Lawrence Summers (1981) and my 1982 article suggest that an increase in anticipated inflation leads to an increase in housing prices, which in turn leads to an increase in construction activity. The analysis presented here suggests that if the government initiates a monetary policy (or any other policy) to stimulate building activity, the policy may actually lead to a decrease in building activity if there is uncertainty about its duration or its effect.

The model also provides insights into the role of real estate speculators who purchase vacant lots, and rather than develop them immediately, choose to keep them vacant for a period of time. By waiting until some future date to build, the speculator is able to construct a building that is most appropriate given economic conditions at that time. Since the exact nature of these economic conditions are unknown at earlier dates, a building constructed earlier will not in general be the optimal size for the future. The decision to build or not build can thus be thought of as weighing the opportunity costs associated with keeping the land vacant against the expected gain from constructing a more appropriate building in the future.

It should also be noted that the framework developed here can easily be extended to analyze other issues relating to real estate pricing under uncertainty. For example, the analysis can easily be augmented to determine the value of houses that may or may not be torn down in the future so that the land can be used to develop large condominium complexes. The framework can also be used to determine when it is optimal to demolish a small building for the purpose of constructing a larger building, and under what conditions it is optimal to renovate an apartment house or convert it to con-

dominiums. One could also use similar techniques to analyze the effect of uncertainty on the optimal durability of buildings.

## REFERENCES

- Bernanke, Ben S., "Irreversibility, Uncertainty, and Cyclical Investment," *Quarterly Journal of Economics*, February 1983, 97, 85-106.
- Black, Fisher and Scholes, Myron, "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, May/June 1973, 81, 637-59.
- Cox, John C., Ross, Stephen A. and Rubinstein, Mark, "Option Pricing: A Simplified Approach," *Journal of Financial Economics*, September 1979, 7, 229-63.
- Cukierman, Alex, "The Effects of Uncertainty on Investment under Risk Neutrality with Endogenous Information," *Journal of Political Economy*, June 1980, 88, 462-75.
- Findlay, M. Chapman and Howson, Hugh R., "Optimal Intertemporal Real Estate Ownership, Valuation, and Use," *American Real Estate and Urban Economics Association Journal*, Summer 1975, 3, 51-66.
- Greenley, Douglas A., Walsh, Richard G. and Young, Robert A., "Option Value: Empirical Evidence from a Case Study of Recreation and Water Quality," *Quarterly Journal of Economics*, November 1981, 95, 657-73.
- Krutilla, John V., "Conservation Reconsidered," *American Economic Review*, September 1967, 57, 777-86.
- Markusen, James and Scheffman, David T., "The Timing of Residential Land Development: A General Equilibrium Approach," *Journal of Urban Economics*, October 1978, 5, 411-24.
- Merton, Robert C., "Theory of Rational Option Pricing," *Bell Journal of Economics*, Spring 1973, 4, 141-83.
- Ratcliff, Richard U., *Valuation for Real Estate Decisions*, Santa Cruz: Democrat Press, 1972.
- Rendleman, Richard J. and Bartter, Brit J., "Two-State Option Pricing," *Journal of Finance*, December 1979, 34, 117-34.
- Shoup, Donald C., "The Optimal Timing of Urban Land Development," *Regional Science Association Papers*, 1970, 25, 33-44.
- Stulz, René M., "Options on the Minimum or

the Maximum of Two Risky Assets: Analysis and Applications," *Journal of Financial Economics*, July 1982, 10, 161-85.

Summers, Lawrence H., "Inflation, The Stock Market, and Owner-Occupied Housing,"

*American Economic Review Proceedings*, May 1981, 71, 429-34.

Titman, Sheridan, "The Effects of Anticipated Inflation on Housing Market Equilibrium," *Journal of Finance*, June 1982, 37, 827-42.

# A Note on Equity and Efficiency in the Pricing of Local Telephone Services

By EDWARD RENSHAW\*

Since the publication of Bridger Mitchell's article on the "Optimal Pricing of Local Telephone Service" (1978), it has been assumed that social welfare can usually be increased by moving from a flat monthly rate for local calls to a two-part tariff with a price per call that is somewhat in excess of marginal cost. While a fixed monthly charge for local calls can be considered a regressive head tax (A. M. Henderson, 1947), it does not follow that a two-part tariff will resolve the equity problem. In this paper I use a simple diagram and a two-person revenue-maximizing formula to illustrate one of the more important limitations of usage-sensitive pricing.

In the following analysis, it is assumed that there are two types of telephone users. The first type of consumer,  $D_1$ , is assumed to have a *net* demand for calls or message units, represented by the linear equation:

$$(1) \quad P = a - bQ.$$

The second consumer,  $D_2$ , is assumed to have a net *indexed* demand of

$$(2) \quad P = 1.0 - Q.$$

I use the term "net demand" to refer to that portion of the consumers demand curve which lies above the marginal cost of providing local telephone service.<sup>1</sup> In the case of

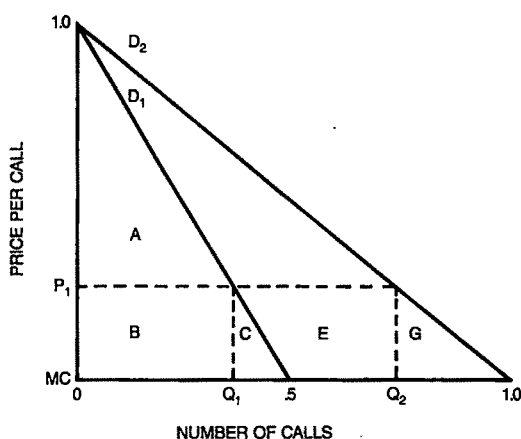


FIGURE 1. INDIVIDUAL DEMAND FUNCTIONS FOR CALLS OF TWO CONSUMERS WITH DIFFERENT TASTE PARAMETERS BUT THE SAME INCOME

the second consumer, the maximum price that can be obtained for the first call or message unit has been arbitrarily indexed to equal 1.0. The number of calls or message units that will take place when the price per unit of service is set equal to marginal cost ( $MC$ ) has also been indexed to equal 1.0.

In Figure 1,  $a$  in equation (1) is set equal to 1.0, and  $b$  equal to 2.0. Note that both of the demand curves have the same price intercept. At any price less than 1.0, however, the second consumer (who had a greater taste for calling) will make twice as many calls as the first consumer.

The maximum monthly flat rate charge that can be collected from  $D_1$  without causing him or her to disconnect will be equal to the consumers' surplus areas under the demand curve  $D_1$  above  $MC$ , or areas  $(A + B$

\*Department of Economics, State University of New York, Albany, NY 12222.

<sup>1</sup>The short-run marginal cost of a local direct dial call is probably nearly equal to zero in many cases. Whether local telephone companies would be motivated to reduce capacity in the long run in response to a usage price in excess of short-run marginal cost is more conjectural and the subject of considerable controversy. In a mobile society it may not be wise to reduce capacity for a resident with little taste for telephoning, since he or she may move and be replaced by someone with a

higher demand for telephone service. For a provocative discussion of marginal cost, see the article by John Wilson (1983).

+  $C$ ) plus some fixed amount ( $F$ ) that represents that person's option demand and willingness to pay for telephone service just to be able to receive calls that are initiated by others.<sup>2</sup>

Suppose that a monthly customer charge equal to ( $F$ ) plus areas ( $A + B + C$ ) is not enough to recover the fixed costs of maintaining local telephone service. Will it be possible to increase telephone revenues by reducing the monthly customer charge to an amount equal to say ( $F$ ) plus area ( $A$ ) and then charge a price per call,  $P_1$ , that is in excess of marginal cost?

A price equal to  $P_1$  will cause the first consumer to cut back his or her usage from .5 to  $Q_1$  and will reduce the total amount of revenue that can be collected from  $D_1$  by an amount equal to area ( $C$ ). Total revenue will increase, however, if area ( $E$ ), the extra revenue to be obtained from the second consumer,  $D_2$ , is greater than area ( $C$ ).

For the demand structure depicted in Figure 1, total revenue can be increased as long as  $P_1$  is less than .25. We can verify this conclusion in a more general way by noting that the total net revenue in excess of marginal cost can be calculated using the following formula:

$$(3) \quad TR = 2F_1 + P_1Q_1 + P_1Q_2 \\ + 2(a - P_1)Q_1(\frac{1}{2}).$$

The expressions  $P_1Q_1$  and  $P_1Q_2$  represent the net usage sensitive revenues to be obtained from  $D_1$  and  $D_2$  when the per unit charge in excess of marginal cost is set equal to  $P_1$ . The expression  $(a - P_1)Q_1(\frac{1}{2})$  is the consumers' surplus area ( $A$ ). This area plus  $F_1$  represent the maximum customer charge that can be collected from  $D_1$  without causing the consumer to go without telephone service. Both of these customer charge elements are multiplied by two to give us the total amount of flat rate revenue to be obtained from  $D_1$  and  $D_2$ .

Equation (3) provides a general definition of the net revenue that will be available to

compensate for fixed costs and is not related in a very specific way to our assumed demand equations.

When equations (1) and (2) are solved from  $Q_1$  and  $Q_2$  and the resulting expressions are substituted into (3), we obtain a more useful definition of the total revenue to be expected in conjunction with a two-part tariff:

$$(4) \quad TR = 2F_1 + (1 - (a/b))P_1 \\ - P^2 + (a^2/b).$$

Taking the derivative of this revenue function with respect to  $P_1$  and setting the resulting expression equal to zero allows one to solve for the usage sensitive price,  $P_1^*$ , which maximizes net telephone revenue

$$(5) \quad P_1^* = \frac{1}{2} - (a/2b).$$

In constructing Figure 1 it was assumed that  $a = 1.0$ , and that  $b = 2.0$ . When these values are substituted into equation (5), we can show that  $P_1^* = .25$ . If both the  $b$  and  $a$  values in equations (1) and (5) were set equal to 1.0, on the other hand, both consumers would have identical demand curves for local calls and the revenue-maximizing price in excess of marginal cost would be  $P_1^* = 0$ . With,  $a = 1.0$ , we can conclude that the higher the  $b$  value, other things equal, the greater the taste differential, and the greater the revenue maximizing price in excess of marginal cost.

Let us now consider a case where  $a$  and  $b$  in equation (1) are both set equal to .5. (See Figure 2.) The maximum customer charge that can be imposed in this case without causing  $D_1$  to disconnect is the same as in Figure 1 if we assume that there is no difference in  $F_1$ , which represents option demand and a willingness to pay for a telephone connection just to be able to receive calls that are initiated by others.

When  $a = b$ , however, one can use equation (5) to quickly discover that a two-part tariff in excess of marginal usage cost will not increase telephone revenue in the two-person or two-group case where  $D_1$  and  $D_2$  type of demanders are equally numerous.

<sup>2</sup>For a more detailed discussion of the demand for telephone service, see the discussion by Lester Taylor (1980).

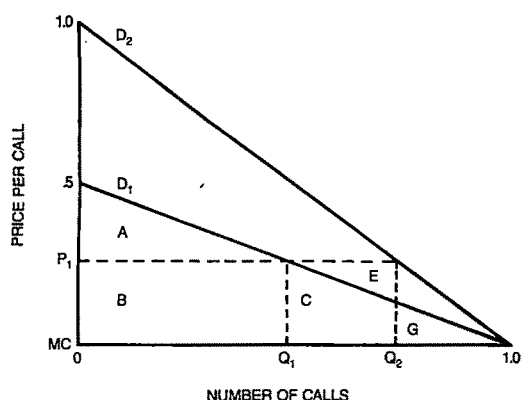


FIGURE 2. INDIVIDUAL DEMAND FUNCTIONS FOR CALLS OF TWO CONSUMERS WITH THE SAME TASTE PARAMETERS BUT DIFFERENT INCOMES

This is a rather startling conclusion that can easily be verified in connection with Figure 2 by noting that if a two-part tariff is to raise more net revenue than a flat rate tariff equal to  $F1$  plus areas  $(A + B + C + G)$ , area  $(E)$  will have to be greater than the consumers' surplus loss areas for both consumers,  $(C + G)$  and  $(G)$ .

In his path-breaking article, Mitchell established the precedent of assuming that persons with similar tastes for telephone service and different incomes will have demand curves with similar call intercepts and different price intercepts.<sup>3</sup> To the extent that this characterization is correct, we can conclude that a two-part tariff will not be successful at equalizing the distribution of effective purchasing power and keeping low-income households connected to the telephone network in an era of deregulation and reduced cross subsidization of local telephone service from fancy equipment rentals and long distance communication.

It follows—if one believes that there are external benefits in having most low-income households connected to the telephone sys-

<sup>3</sup> Mitchell's further assumption that there is quite a bit of income variation as well as taste variation no doubt helps to explain why his optimal prices for local telephone service were not very much in excess of his assumed marginal cost.

tem—that serious consideration should be given to other alternatives for furthering the goal of universal service such as well targeted subsidies or rebates to low-income households and geographical price discrimination which would require persons in high-income neighborhoods to pay more for basic telephone service than persons living in low-income neighborhoods.<sup>4</sup>

Census data indicate, in any event, that a considerable part of the variation in local telephone connections can be explained on the basis of income differences. In Nassau County, New York, where the median household income was \$26,091 in 1980, less than 2 percent of the surveyed households were without a telephone. In the not very distant Bronx, on the other hand, where median household income was only \$10,947, more than 17 percent of the households were without a telephone. The best predictor of the proportion of households without a telephone for New York's 62 counties, however, is not median income but the proportion of households below the poverty line.<sup>5</sup>

<sup>4</sup> Geographical price discrimination on the basis of income is somewhat analogous to financing part of the fixed costs of local telephone service with a property tax. In many instances it can be defended on the basis of extra costs of providing telephone service in less thickly settled affluent areas.

<sup>5</sup> The  $r^2$  for the proportion of households below the poverty line is .747 and can be compared to  $r^2$ 's of .554 for median family income and .429 for mean income.

## REFERENCES

- Feldstein, Martin, "Equity and Efficiency in Public Sector Pricing: The Optimal Two-Part Tariff," *Quarterly Journal of Economics*, May 1972, 86, 175-87.
- Henderson, A. M., "The Pricing of Public Utility Undertakings," *Manchester School*, September 1947, 15, 223-50.
- Mitchell, Bridger, "Optimal Pricing of Local Telephone Service," *American Economic Review*, September 1978, 68, 517-37.
- Renshaw, Edward, "On the Distribution of Telephone Communication Subsidies,"

*Public Utilities Fortnightly*, July 21, 1983, 112, 34-39.

Taylor, Lester D., "The Demand for Telecommunications: A Nontechnical Exposition," in Michael Crew, ed., *Issues in Public-Util-*

*ity Pricing and Regulation*, Lexington: Lexington Books, 1980, ch. 6.

Wilson, John, "Telephone Access Costs and Rates," *Public Utilities Fortnightly*, September 15, 1983, 112, 18-25.



# Women, Work, and Divorce

By WILLIAM SANDER\*

An upward trend in the divorce rate in the U.S. has been attributed to an increase in the earning ability of women (Gary Becker et al., 1977, and Becker, 1981). The reasoning for the increase in marital instability is that high-wage (quality) women "gain less from marriage than other women do because higher earnings reduce the demand for children and the advantages of the sexual division of labor in marriage" (Becker, p. 231).

To some extent, an increase in the earning ability of women is a consequence of a higher divorce rate (Becker). That is, divorced women tend to invest more in work experience and thus command a higher wage.

The economic explanation for a higher divorce rate has been questioned by some (for example, Mark Blaug, 1980) on the grounds that changes in the family structure over time reflect changes in tastes rather than exogenously determined economic effects. That is, investments in work by women reflect, perhaps, an increase in a taste for work and a decrease in a taste for marriage.

In this paper, I present evidence that adds support to the economic approach to divorce. I will show that the divorce rate is significantly and substantially affected by the earning ability of women in market work. Data will be drawn from the U.S. farm sector. After a brief overview of the economic theory of divorce and its relevance to the farm sector, data sources and regression specifications will be discussed.

## I. Theory

A complete treatment of the economic theory of divorce is beyond the scope of this

paper (see Becker et al., and Becker). The most important aspect of the theory, though, is that high-wage women gain less from marriage relative to other women because the gains from specialization within marriage (the wife in household work and the husband in market work) are less. Since high-wage women tend to marry high-wage men, the costs of low-level specialization within the household are at least partly offset by the benefits to women of a high-wage husband. Thus, the net effect would seemingly be ambiguous.

Perhaps more importantly, the shadow price of divorce is less for high-wage women. The reason for this is simply that the marginal economic value of a loss in husband's income is less for a high-wage wife. That is, an increase in the earning ability of women enables them to leave an unhappy marriage and either remain divorced or remarry.

An increase in the male wage should have a positive effect on the gains from marriage because high-wage men tend to attract high-wage wives. For this reason, the independent effect of the male wage on the divorce rate has been shown to be negative.

With respect to the farm sector, the farm divorce rate has been substantially below the urban and rural nonfarm divorce rates (Table 1). In addition, the percent of farm and rural nonfarm men who have been divorced exceeds the rate for farm and rural nonfarm women. On the other hand, there are more divorced women living in urban areas than men. This probably reflects the economic attractiveness of urban areas for divorced women relative to divorced men. That is, divorced women would tend to gain more from migration relative to divorced men because the latter would already tend to be located where their earning ability was relatively high.

The economic theory of divorce indicates that farm families should experience less marital instability than their nonfarm counterparts. Two important reasons for this are that farm wives specialize more in household

\*Assistant Professor, Department of Family and Consumer Economics, University of Illinois at Urbana-Champaign, Urbana, IL 61801. I thank Theodore W. Schultz, Gary S. Becker, and the participants of the agricultural economics workshop at Chicago for comments on a preliminary draft.

TABLE 1—PERCENT OF MEN AND WOMEN EVER MARRIED WHO WERE KNOWN TO HAVE BEEN DIVORCED, BY LOCATION, 1970

Location	Divorce Rates	
	Male	Female
Urban	14.6	15.5
Rural Nonfarm	13.1	12.0
Rural Farm	7.4	6.6

Source: U.S. Department of Commerce, 1973.

(including on-farm) work and less in off-farm market work relative to their nonfarm counterparts. In 1979, 44 percent of farm women were labor force participants. About one-third of these were counted in the labor force via their farm work, while two-thirds were in the labor force via off-farm market work (U.S. Department of Labor, 1980, and Howard Hayghe, 1981). About half of their nonfarm counterparts were in the labor force at this time.

Farm women are more specialized in farm work than the labor force participation statistics indicate. A recent national survey of farm women indicates that most farm women participate in some farm tasks (National Opinion Research Center, 1980). Thus, there is a greater sexual division of labor in the farm household relative to the nonfarm household. It should follow that this increases the gains from marriage for farm wives relative to other wives, to the extent that specialization appreciably affects the gains from marriage.

The earning ability of farm men has increased over time. T. W. Schultz<sup>1</sup> estimates that the average value of the farmer's time is now about the same as the average value of time in the nonfarm sector. I would note that this should be particularly true regarding comparisons of the farm and rural nonfarm sectors. That is, farmers in a state have about the same earning ability as their rural nonfarm counterparts because the two sectors are highly integrated. Therefore, the differential in the earning ability between farm and

nonfarm men cannot be drawn upon to explain a low farm divorce rate. It will be demonstrated that the relatively low level of specialization by farm women in market work explains a substantial part of the farm-rural nonfarm divorce gap.

## II. Data and Specifications

The male farm divorce rate (percent ever been divorced) will be estimated below. The male rate rather than the female rate will be used because it better reflects the real farm rate. As noted above, divorced farm women probably migrate at a higher rate from agriculture than divorced farm men. There is no evidence available regarding migration of divorced farm men from farming.

The key variables that would affect the gains from marriage and, accordingly, the costs of divorce are the husband's wage and the value of the wife's time in household (including on-farm) and market work. In addition, it has been shown that religion has a significant and substantial effect on the cost of divorce.

Data are not available to estimate the independent effects of these variables. However, useful proxies are available that enable one to explain a good part of the pattern in the farm divorce rate. Since the farm and rural nonfarm economies are highly integrated, particularly at the state level, the rural nonfarm divorce rate should pick up the net effect of many of the determinants of the farm divorce rate in a state. In particular, it should pick up the effects of the earning ability of farm men, the level of schooling in the farm sector, and variables that affect the cost of divorce such as religion. To some extent, the correlation coefficients between farm and rural nonfarm average family income in 1970 (.95) and farm and rural nonfarm schooling (.93) indicate the former. No data are available on religion at the substate level.

Since it is not unrealistic to assume that the farm and rural nonfarm economies are highly integrated in a state, it should be expected that the correlation coefficient between the farm and rural nonfarm divorce rate should be high, as well. It was .91 for farm and rural nonfarm men in 1970.

<sup>1</sup>Personal communication, 1984.

However, there is some variability in the difference between the average farm and rural nonfarm divorce rate for men by state. In 1970, the standard deviation for this difference was 2.1. The mean difference was 5.7 in 1970 (see Table 1).

I believe that the difference and the variability in the difference between the farm and rural nonfarm divorce rates for men is primarily a product of differences in the earning ability of farm wives in market work relative to rural nonfarm wives. This difference is indicated to some extent by the relatively low correlation coefficient between the labor force participation rate for farm women and rural nonfarm women. (.55 in 1970). It is important to note that farm women had about the same quantity of schooling as their rural nonfarm counterparts in 1970, although urban women had more schooling. Thus, the level of schooling of farm wives is not the reason for their relatively low level of participation in off-farm market work. That is, the reason that farm women do not acquire market-specific capital relative to their rural nonfarm counterparts is not a product of less schooling.

Although many factors affect labor force participation rates, with state-level data population density was found to be the only significant determinant of the labor force participation rate for farm women and will be used in regressions to reflect this. This should be expected since most farm women are labor force participants via off-farm market work and since the distance to an off-farm job would tend to be less in more densely populated states. That is, the effective off-farm wage rate would tend to be higher for farm women in more densely populated states because location is a key determinant of the acquisition of market-specific capital by farm women. Population density is an interesting measure of the economic value of a woman's time in market work since it is highly exogenous. That is, density is not a consequence of a high divorce rate for farm men in a state.

In addition, the farm-rural nonfarm divorce rate differential could be related to the greater sexual division of labor within the farm household. That is, the gains from marriage might be higher for farm wives if they

TABLE 2—SUMMARY STATISTICS FOR DATA SET, 1970

Variable	Mean <sup>a</sup>	Standard Deviation
Farm Divorce Rate	8.6 <sup>b</sup>	3.7
Rural Nonfarm Divorce Rate	13.8 <sup>b</sup>	4.8
Farm Assets	\$94,511	72,431
State Population Density	143 <sup>c</sup>	222
Labor Force Participation, Farm Women	29.7 <sup>b</sup>	5.4
Divorce Rate Differential	5.2 <sup>b</sup>	2.1

<sup>a</sup>Unweighted

<sup>b</sup>Shown in percent.

<sup>c</sup>People in one square mile.

had acquired more marital-specific capital relative to their rural nonfarm counterparts. Without a doubt, farm wives acquired relatively more specific capital. Whether this has a significant effect upon the gains from marriage remains an empirical issue. Farm assets are undoubtedly related to the acquisition of specific capital by wives and will be used in my farm divorce estimates below. Farm assets may also partly reflect the wealth (gains from marriage) that farm wives acquire upon entry into marriage. Thus, farm assets may reflect gains acquired via wealth and specialization.

To sum up, the explanatory variables that will be used to explain variations in the percent of farm men in a state who were known to have been divorced at some time are the rural nonfarm divorce rate for men, farm assets, and population density. In addition, the labor force participation rate for farm women will be added in some regressions to show its effect. As noted above, this variable is a more biased measure of the value of a wife's time in work. Summary statistics are presented in Table 2.

### III. Results

The results of ordinary least squares (OLS) estimates of the "farm divorce rate" (Table 3) are as follows. The rural nonfarm divorce rate had a highly significant and substantial positive effect on the farm divorce rate as one would expect. That is, the rural nonfarm divorce rate is correlated with the farm divorce rate at the state level because some of the determinants of divorce are correlated at

TABLE 3—OLS ESTIMATES OF THE FARM DIVORCE RATES, 1970

	(1)	(2)	(3)	(4)
Population Density	.003 <sup>a</sup> (3.3)	.004 <sup>a</sup> (4.3)	.004 <sup>a</sup> (4.2)	
Labor Force Participation, Farm Women	.06 (1.5)			.12 <sup>a</sup> (3.0)
Farm Assets	-.002 (.6)	-.004 (1.3)		
Rural Nonfarm Divorce Rate	.76 <sup>a</sup> (13.5)	.79 <sup>a</sup> (14.8)	.75 <sup>a</sup> (17.4)	.69 <sup>a</sup> (15.6)
Constant	-4.1	-2.5	-2.3	-4.5
R <sup>2</sup>	.87	.87	.87	.85

Note: *t*-statistics are shown in parentheses.

<sup>a</sup>significant at the 5 percent level.

the rural nonfarm and farm levels, as well. Farm assets had a negative sign as would be expected; however, the coefficient was not statistically significant. Population density had a significant positive effect. If population density entered the equation, the labor force participation variable for farm women had a positive sign although it was not significant. This variable was shown to have a significantly positive effect if population density was excluded in the regression. In preliminary work with the data, I could not show that any other variable had a statistically significant effect on the farm-rural nonfarm divorce differential, or any significant effect on my results.

My results provide support for the hypothesis that the divorce rate is positively related to the earning ability of women in market work. In addition, if the regression coefficient for density takes the value of .004 the elasticity at means of density and the divorce rate is .07. This implies that if density took on the value of the most densely populated state, it would increase the farm divorce rate by 40 percent over its mean value. Finally, I regressed the difference between the rural nonfarm divorce rate (for men) and the farm divorce rate (state-level data) on density and average farm assets. The result was:

$$\text{Divorce Differential} = 4.8 + .01 \text{ Assets} - .005 \text{ Density.}$$

(3.9)                      (5.0)

The adjusted  $R^2$  was .47 (*t*-statistics are shown in parentheses). In this regression, I show that an increase in farm assets increases the difference between the rural nonfarm and farm divorce rates. This is the expected result since the gains from marriage should be higher on larger farms. An increase in population density reduces the difference which is consistent with the results in Table 3.

#### IV. Conclusions

It has been shown that an exogenous increase in the earning ability of women substantially increases the divorce rate. That is, in states where farm women have a higher incentive to acquire market specific capital, it becomes more likely that at some time they will become divorced.

The hypothesis that population density affects the farm-rural nonfarm divorce differential via the incentive to invest in market work seems to follow the data. In 1970, the difference between the rural nonfarm and farm divorce rates (percent ever divorced) was the greatest in three sparsely populated states—Nevada, Wyoming, and Arizona. The difference was minimized in three densely populated states—New Jersey, Connecticut, and Massachusetts.

The results regarding the effect of farm assets on the gains from marriage were less illuminating. In estimating the farm divorce rate, farm assets did not seem to be consequential. However, assets did seem to increase the farm-rural nonfarm divorce differential. Thus, less can be said of the effect of assets on the gains from marriage.

However, the divorce rate differential tended to be relatively high in the North Central states where farm assets tend to be high. Included in this group are Illinois, Indiana, Ohio, and Michigan. In these states, the opportunity cost of investments in market work experience by farm wives tend to be high because of the relatively high value of the farm wife's time in on-farm work.

#### REFERENCES

- Becker, Gary S., *A Treatise on the Family*, Cambridge: Harvard University Press,

- 1981.
- \_\_\_\_\_, Landes, Elisabeth M. and Michael, Robert T., "An Economic Analysis of Marital Instability," *Journal of Political Economy*, December 1977, 85, 1141-87.
- Blaug, Mark, *The Methodology of Economics*, Cambridge: Cambridge University Press, 1980.
- Hayghe, Howard, "Husbands and Wives as Earners: An Analysis of Family Data," *Monthly Labor Review*, February 1981, 46-59.
- National Opinion Research Center, "Survey of Farm Women," Chicago, 1980.
- U.S. Department of Commerce, *Census of Population: 1970*, Washington: USGPO, 1973.
- \_\_\_\_\_, *Statistical Abstract of the United States: 1979*, Washington: USGPO, 1980.
- U.S. Department of Labor, *Handbook of Labor Statistics*, Washington: USGPO, 1980.

# Money, Anticipated Changes, and Policy Effectiveness

By RICHARD G. SHEEHAN\*

The Lucas-Sargent-Wallace (L-S-W) proposition states that expected policy changes have no influence on real economic variables even in the short run. (See Robert Lucas, 1972, and Thomas Sargent and Neil Wallace, 1975.) One form of empirical test of the L-S-W proposition (the neutrality proposition) began with Robert Barro's (1977; 1978) seminal articles supporting the neutrality proposition. However, more recent papers, including those by David Small (1979), John Makin (1982), and Frederic Mishkin (1982a, b) have reexamined the relationships between anticipated and unanticipated money vs. output, unemployment, and inflation, and have found evidence suggesting that anticipated monetary changes do have real effects. All of the above-cited studies use expectations procedures which estimate, for example, a money growth equation over the entire period under consideration.<sup>1</sup> However, this approach assumes that when expectations are being formed at the beginning of the period, economic agents have the relevant information for the entire period. In this note I reexamine the impact of expected vs. unexpected monetary changes focusing on the informational restrictions actually

faced by economic agents when formulating expectations.

The structure of the model used here is similar to the models of Barro and Mishkin. A measure of real output is assumed to be a function of a policy variable and a vector of predetermined variables. The policy variable can be decomposed into anticipated and unanticipated components. Algebraically:

$$(1) \quad \dot{y}_t = \alpha_1 Z_{1t} + \beta_1 (\dot{m}_t - {}_{t-1}\dot{m}_t) + \delta_1 \dot{m}_t + \varepsilon_{1t}$$

$$(2) \quad \dot{m}_t = \alpha_2 Z_{2t} + \varepsilon_{2t}$$

$$(3) \quad \Sigma = E((\varepsilon_{1t} \varepsilon_{2t})' (\varepsilon_{1t} \varepsilon_{2t})) = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

where  $\dot{y}$  is the rate of change in a measure of real output,  $Z_1$  and  $Z_2$  are vectors of predetermined variables,  $\dot{m}_t$  is the actual rate of change in a policy variable such as the money supply, and  ${}_{t-1}\dot{m}_t$  refers to the expected rate of change of the policy variable at time  $t$ , conditional on the information available at time  $t-1$ . The  $\varepsilon_{1t}$  and  $\varepsilon_{2t}$  are serially uncorrelated normally distributed error terms, distributed as  $\eta(0, \sigma_1^2)$  and  $\eta(0, \sigma_2^2)$ , respectively. It is further assumed they are contemporaneously independent. Assuming economic agents know the model and form econometrically rational expectations, then

$$(4) \quad {}_{t-1}\dot{m}_t = \alpha_{2t-1} Z_{2t}$$

It is further assumed that economic agents have limited information. For example, economic agents in 1964:1 only have the data actually published before 1964:1. A data revision in, say, 1968:2 for the money stock in 1963:4 does not become available until 1968:3. Each period new data becomes available and any old data may be revised. Both changes may be important when generating expectations. Not only may the estimated

\*Research and Public Information Department, Federal Reserve Bank of St. Louis, P.O. Box 442, St. Louis, MO 63166. I am grateful to Robin Grieves, William Hosek, David Schirm, Geoffrey Woglom, Frank Zahn, and an anonymous referee for comments on an earlier draft. Any errors remaining are, of course, entirely my responsibility. The views and opinions expressed herein are solely my own and do not necessarily represent those of the Federal Reserve Bank of St. Louis or the Federal Reserve System.

<sup>1</sup>Barro (1977) states that he also considered money growth expectations based solely on prior observations but found little change from his reported results. Steven Sheffrin (1979), using an ARIMA model, and Jacob Grossman (1979), using weighted least squares, also add data period by period. However, all three of these studies apparently use the most recent data with all revisions and not the data actually available to economic agents at some earlier point in time.

money growth equation change from period to period, the data used to estimate that equation may also change from period to period. Monetary surprises may occur because of random fluctuations, exogenous shocks, changes in the monetary authority's behavior, or data revisions.

Alternate money forecasting equations are used to test the sensitivity of the results to the method for determining expectations. The forecasting equations differ partly in the choice of included variables. Following Barro, the equation (2) is viewed as a reaction function where monetary changes depend in part on what macro variable the monetary authorities are attempting to control. Algebraically equation (2) can be rewritten as

$$(5) \quad \dot{m}_t = \alpha_3 Z_{3t} + \beta_3 (\dot{X}_{t-1} - \dot{X}_{t-2}) + \varepsilon_{3t},$$

which implies that monetary authorities are attempting to control the expected rate of change of a goal variable  $X$ . Equation (5) is repeatedly reestimated each time using one additional period of data together with any data revisions made in the previous quarter. Anticipated and unanticipated monetary changes are obtained by sequentially forecasting money growth one period ahead using the reestimated equation (5). This procedure is in keeping with the process of letting economic agents revise their estimated forecasting equations only as data actually become available.<sup>2</sup>

<sup>2</sup>Some assumptions must be made concerning when the forecast for a quarter is made and what information is then available. These assumptions are, necessarily, somewhat arbitrary. We assume that economic agents when forecasting, say, 1964:1 have information available in issues of *Business Conditions Digest* up to and including November 1963. Similarly, when forecasting 1964:2 values, economic agents have information from issues of *BCD* prior to and including February 1963. Information is likewise limited throughout the sample. This approach assumes that data revisions at all times have a zero expected mean. If the mean is not zero, then economic agents would anticipate some revision and would incorporate that in their forecasts. The anticipated data revision would then be of concern.

The money forecasting equations are estimated using *OLS* with the dependent specified as the difference in the log of the money stock ( $M1$ ). The equations include as independent variables a constant, a time trend, quarterly dummy variables, eight lags of the dependent variable, and eight lags (of the difference in the log) of either the unemployment rate (to generate  $ME1$ ), the *GNP* deflator (for  $ME2$ ), or nominal *GNP* (for  $ME3$ ). The former is included to facilitate comparison with Barro's original studies, while the latter two were included to allow monetary policymakers to target variables over which they should have some control even under the neutrality proposition. A fourth expectations series ( $ME4$ ) is generated using Jacob Grossman's method of weighted least squares. It includes lagged nominal *GNP* as a dependent variable, and uses the most recent forty observations available at a particular point in time.<sup>3</sup> The first four expectations series on a period-by-period basis reestimate money growth, forecast one period ahead, and use that forecast to generate expected and unexpected changes during the period for which the equation was estimated. A final expectations series ( $ME5$ ) is generated using Barro's basic approach. Anticipated money is estimated using the data available only at the end of the period with the unemployment rate as an explanatory variable.

"Second-stage" equations for real output, the unemployment rate, and the inflation rate are then estimated. These equations include (for the vector  $Z_1$ ) a constant, a time trend, seasonal dummy variables, and eight lags of the dependent variables. Fourth-order polynomial distributed lags (*PDL*) are used for expected ( $\dot{m}^e$ ) and/or unexpected ( $\dot{m}^u$ ) differences in the log of the money stock. The *PDL* specification is chosen, following Mishkin, because it allows long lags on the

<sup>3</sup>Economic agents may rationally believe more recent observations are more important in ascertaining future values. However, any attempt to weight recent observations more heavily would appear to face an insurmountable obstacle in generating a nonarbitrary set of weights. The weights used here are the same as those of Grossman.

TABLE 1— $\chi^2$  RESULTS

	Constraint	ME1	ME2	ME3	ME4	ME5
<b>Short Lags</b>						
Real Output	$\alpha = 0$ given $\beta = 0$	34.63 <sup>a</sup>	41.90 <sup>a</sup>	36.76 <sup>a</sup>	11.68 <sup>b</sup>	42.37 <sup>a</sup>
	$\beta = 0$ given $\alpha = 0$	2.21	10.97 <sup>b</sup>	3.85	5.14	38.06 <sup>a</sup>
	$\beta = 0$ given $\alpha \neq 0$	13.37 <sup>a</sup>	11.52 <sup>b</sup>	11.53 <sup>b</sup>	37.78 <sup>a</sup>	6.64
Unemployment Rate	$\alpha = 0$ given $\beta = 0$	26.43 <sup>a</sup>	28.71 <sup>a</sup>	27.03 <sup>a</sup>	9.46 <sup>c</sup>	35.59 <sup>a</sup>
	$\beta = 0$ given $\alpha = 0$	1.89	7.61	3.90	4.78	22.79 <sup>a</sup>
	$\beta = 0$ given $\alpha \neq 0$	4.12	6.08	4.70	21.89 <sup>a</sup>	4.21
Prices	$\alpha = 0$ given $\beta = 0$	17.25 <sup>a</sup>	17.97 <sup>a</sup>	17.05 <sup>a</sup>	7.79 <sup>c</sup>	13.36 <sup>a</sup>
	$\beta = 0$ given $\alpha = 0$	2.25	3.82	1.39	4.07	16.79 <sup>a</sup>
	$\beta = 0$ given $\alpha \neq 0$	4.80	1.84	2.84	13.90 <sup>a</sup>	16.96 <sup>a</sup>
<b>Long Lags</b>						
Real Output	$\alpha = 0$ given $\beta = 0$	32.23 <sup>a</sup>	34.82 <sup>a</sup>	33.10 <sup>a</sup>	15.82 <sup>a</sup>	31.45 <sup>a</sup>
	$\beta = 0$ given $\alpha = 0$	2.07	6.72	2.09	5.19	32.86 <sup>a</sup>
	$\beta = 0$ given $\alpha \neq 0$	16.12 <sup>a</sup>	14.71 <sup>a</sup>	16.70 <sup>a</sup>	26.39 <sup>a</sup>	13.87 <sup>a</sup>
Unemployment Rate	$\alpha = 0$ given $\beta = 0$	26.43 <sup>a</sup>	28.89 <sup>a</sup>	26.82 <sup>a</sup>	14.40 <sup>a</sup>	27.90 <sup>a</sup>
	$\beta = 0$ given $\alpha = 0$	11.29 <sup>b</sup>	9.35 <sup>c</sup>	3.24	8.60 <sup>c</sup>	23.02 <sup>a</sup>
	$\beta = 0$ given $\alpha \neq 0$	15.82 <sup>a</sup>	17.49 <sup>a</sup>	17.31 <sup>a</sup>	23.88 <sup>a</sup>	7.61
Prices	$\alpha = 0$ given $\beta = 0$	19.86 <sup>a</sup>	20.89 <sup>a</sup>	19.36 <sup>a</sup>	17.23 <sup>a</sup>	19.46 <sup>a</sup>
	$\beta = 0$ given $\alpha = 0$	1.34	4.82	.84	3.76	20.74 <sup>a</sup>
	$\beta = 0$ given $\alpha \neq 0$	20.45 <sup>a</sup>	19.68 <sup>a</sup>	27.32 <sup>a</sup>	22.19 <sup>a</sup>	28.71 <sup>a</sup>

Note: The basic equation is  $\dot{y} = \alpha \dot{m} + \beta \dot{m}e + \Psi Z_4 + e$ , where  $Z_4$  is a vector which includes a constant, a time trend, quarterly dummy variables, and eight lags of the dependent variable.

<sup>a</sup>Significant at the 99 percent level

<sup>b</sup>Significant at the 95 percent level

<sup>c</sup>Significant at the 90 percent level

independent variable while conserving degrees of freedom. Following Mishkin, both short- and long-lag specifications of the *PDL* are used, the short-lag specifications including eight lags and the long-lag results including twenty lags.<sup>4</sup> Comparing the log-likelihood ratios when the monetary changes are included vs. excluded allows us to test whether the coefficients on either set of terms should be constrained to equal zero. The test statistic is 2 (log likelihood of included—log likelihood of excluded) which is distributed  $\chi^2$  with degrees of freedom equal to the number of zero restrictions. (Since I use fourth-order *PDLs* for both the  $\dot{m}$  and  $\dot{m}e$  terms, the number of zero restrictions is four per constrained series.) Table 1 presents the log-likelihood comparisons for all second-stage equations with real output, the unem-

ployment rate, and inflation as the dependent variables.<sup>5</sup>

Table 1 presents  $\chi^2$  statistics testing the joint significance of all anticipated or unanticipated money changes using alternate specifications of the money reaction function. The short- and long-lag results in Table 1 are unanimous in their support for the contention that unanticipated policy changes have a significant impact on real output, the unemployment rate, and inflation. However, there are considerable differences in the anticipated change variables. In the short-lag equations, most limited information procedures (*ME1*–*ME3*) suggest that expected money changes influence real output but not the unemployment rate or inflation. The weighted least squares approach (*ME4*) indicates expected policy changes influence all

<sup>4</sup>These lag structures are employed simply to facilitate comparison of these results with those of Mishkin. No attempt was made to search for the "preferred" *PDL* specification.

<sup>5</sup>The money forecasting equations are originally estimated from 1954:1 to 1965:4 and are updated through 1981:3. The second-stage equations are then estimated from 1966:1 to 1981:4.



variables. By contrast, Barro's approach (*ME5*) indicates expected policy changes only influence inflation—a result consistent both with Barro's original conclusion and with the neutrality proposition. Only *ME5* with its extreme informational availability assumption is consistent with the neutrality proposition. The results for *ME5* are also consistent with the hypothesis that the expectations are "too good."<sup>6</sup>

The long-lag results also generally provide little comfort for proponents of the neutrality proposition. When the economic agents are restricted in the information to which they have access, then the long-lag results are unanimous in their declaration that expected policy changes influence measures of real economic activity as well as prices. Only in the extreme information case (*ME5*) is the neutrality proposition even partially supported with the conclusion that expected monetary changes do not influence the unemployment rate but do influence prices (and real output).

Table 2 presents the estimated coefficients on *me* and *mu* for one specification of the reaction function, *ME3*. This specification is chosen because it assumes that information is gradually made available to economic agents, because it assumes that economic agents believe that the Federal Reserve targets a variable which the neutrality proposition states it can influence (inflation), because it imposes no a priori weights on prior information used, and because it generally yields the highest log-likelihood ratios. The coefficients suggest, not surprisingly, that both expected and unexpected money changes will initially increase real output and lower the unemployment rate although those changes are largely offset in the long run. The results also suggest that anticipated

changes in the money stock may have even larger short-run impacts than unanticipated changes. The estimated coefficients also suggest that it will take in the vicinity of one year for the inflationary impact on monetary policy to begin to be felt. Finally, it should be noted that the sum of the estimated coefficients on both expected and unexpected money changes are significantly different from one—the value predicted by the neutrality proposition.

The results presented here strongly suggest that the neutrality proposition does not hold in the short run. Anticipated and unanticipated monetary changes both have impacts on real economic variables as well as prices when information is only gradually made available to economic agents. As has been pointed out by Alan Blinder and Robert Gordon in their discussion immediately following Barro and Mark Rush's paper (1980), the neutrality proposition depends primarily on the assumption of flexible prices. Thus these results indirectly suggest prices are not completely flexible in the short run.<sup>7</sup>

My results also suggest that the decision to accept or reject the neutrality proposition is sensitive to the technique used to generate expectations. Specifically, making extreme informational assumptions about the availability of data would make one more likely to accept the neutrality proposition. By contrast, when data is only gradually revealed to economic agents and those agents gradually revise their forecasting procedure, then anticipated monetary variations are no longer neutral in the short run.

<sup>6</sup>Makin has pointed out that if the neutrality hypothesis is true and if some unanticipated policy changes are erroneously labeled anticipated, then the coefficients on the anticipated changes would be biased away from zero. Thus incorrect modeling of expectations could lead to a false rejection on the neutrality proposition. The *ME5* results for  $\beta = 0$  given  $\alpha = 0$  strongly suggest this possibility.

<sup>7</sup>The neutrality proposition also depends on rational expectations. I do not attempt to test the hypothesis that expectations are formed rationally. It is taken as a maintained hypothesis. Given that expectations are in general unobservable, with differences possible between individuals, and potentially subject to nearly continuous change, I do not feel the assumption of rational expectations can be meaningfully tested in the context of this paper. Of course, the results here, as all previous results rejecting the neutrality proposition, can be alternately interpreted as rejection of the hypothesis that the expectations formation process is accurately modeled. However, that interpretation is made somewhat more difficult by the agreement of the alternate forms *ME1*–*ME4*.

TABLE 2—EXPECTATIONS: ME3

	Real Output		Unemployment Rate		Prices	
<i>mu</i>	.435 <sup>a</sup>	.457 <sup>a</sup>	-1.816 <sup>b</sup>	-1.644 <sup>b</sup>	-.160 <sup>a</sup>	-.052
<i>mu</i> <sub>-1</sub>	.468 <sup>a</sup>	.419 <sup>a</sup>	-1.535 <sup>a</sup>	-1.871 <sup>a</sup>	-.092 <sup>b</sup>	-.072 <sup>b</sup>
<i>mu</i> <sub>-2</sub>	.413 <sup>a</sup>	.359 <sup>a</sup>	-1.269 <sup>b</sup>	-1.893 <sup>a</sup>	-.031	-.064 <sup>b</sup>
<i>mu</i> <sub>-3</sub>	.312 <sup>a</sup>	.286 <sup>a</sup>	-.996 <sup>c</sup>	-1.759 <sup>a</sup>	-.027	-.036
<i>mu</i> <sub>-4</sub>	.200 <sup>c</sup>	.205 <sup>a</sup>	-.709 <sup>c</sup>	-1.512 <sup>a</sup>	-.015	.005
<i>mu</i> <sub>-5</sub>	.102	.122 <sup>c</sup>	-.424	-1.193 <sup>a</sup>	-.007	.053 <sup>b</sup>
<i>mu</i> <sub>-6</sub>	.034	.042	-.173	-.836 <sup>b</sup>	-.002	.103 <sup>a</sup>
<i>mu</i> <sub>-7</sub>	.001	-.032	-.008	-.472 <sup>b</sup>	.001	.150 <sup>a</sup>
<i>mu</i> <sub>-8</sub>	-	-.098	-	-.125	-	.191 <sup>a</sup>
<i>mu</i> <sub>-9</sub>	-	-.148	-	.185	-	.223 <sup>a</sup>
<i>mu</i> <sub>-10</sub>	-	-.186	-	.441	-	.245 <sup>a</sup>
<i>mu</i> <sub>-11</sub>	-	-.210	-	.663	-	.254 <sup>a</sup>
<i>mu</i> <sub>-12</sub>	-	-.219	-	.754	-	.251 <sup>a</sup>
<i>mu</i> <sub>-13</sub>	-	-.214	-	.803	-	.236 <sup>a</sup>
<i>mu</i> <sub>-14</sub>	-	-.196 <sup>c</sup>	-	.783	-	.210 <sup>a</sup>
<i>mu</i> <sub>-15</sub>	-	-.168 <sup>c</sup>	-	.702	-	.176 <sup>a</sup>
<i>mu</i> <sub>-16</sub>	-	-.132 <sup>c</sup>	-	.574	-	.135 <sup>a</sup>
<i>mu</i> <sub>-17</sub>	-	-.093 <sup>b</sup>	-	.416	-	.092 <sup>a</sup>
<i>mu</i> <sub>-18</sub>	-	-.055 <sup>c</sup>	-	.250	-	.052 <sup>a</sup>
<i>mu</i> <sub>-19</sub>	-	-.021	-	.103	-	.019 <sup>a</sup>
<i>me</i>	.787	3.379 <sup>a</sup>	-8.108	-15.670 <sup>b</sup>	.431	-.842 <sup>b</sup>
<i>me</i> <sub>-1</sub>	.962	2.251 <sup>a</sup>	-3.601 <sup>a</sup>	-10.750 <sup>a</sup>	.037	-.814 <sup>a</sup>
<i>me</i> <sub>-2</sub>	.645	1.322 <sup>a</sup>	.350	-6.862 <sup>a</sup>	-.261	-.637 <sup>a</sup>
<i>me</i> <sub>-3</sub>	.223	.577	2.579	-3.868 <sup>c</sup>	-.387	-.365 <sup>b</sup>
<i>me</i> <sub>-4</sub>	-.029	.002	2.676	-1.635	-.326	-.046
<i>me</i> <sub>-5</sub>	-.046	-.419	.986	-.043	-.122	.279
<i>me</i> <sub>-6</sub>	.121	-.700	-1.388	1.019	.121	.575 <sup>a</sup>
<i>me</i> <sub>-7</sub>	.258	-.858 <sup>c</sup>	-2.588	1.655	.236	.815 <sup>a</sup>
<i>me</i> <sub>-8</sub>	-	-.908 <sup>c</sup>	-	1.955	-	.979 <sup>a</sup>
<i>me</i> <sub>-9</sub>	-	-.868 <sup>b</sup>	-	2.002	-	1.055 <sup>a</sup>
<i>me</i> <sub>-10</sub>	-	-.754 <sup>c</sup>	-	1.871	-	1.037 <sup>a</sup>
<i>me</i> <sub>-11</sub>	-	-.584 <sup>c</sup>	-	1.624	-	.925 <sup>a</sup>
<i>me</i> <sub>-12</sub>	-	-.376	-	1.316	-	.728 <sup>a</sup>
<i>me</i> <sub>-13</sub>	-	-.149	-	.992	-	.462 <sup>a</sup>
<i>me</i> <sub>-14</sub>	-	-.079	-	.686	-	.149
<i>me</i> <sub>-15</sub>	-	.288	-	.424	-	-.182 <sup>b</sup>
<i>me</i> <sub>-16</sub>	-	.458	-	.221	-	-.495 <sup>a</sup>
<i>me</i> <sub>-17</sub>	-	.648 <sup>c</sup>	-	.085	-	-.744 <sup>a</sup>
<i>me</i> <sub>-18</sub>	-	.600	-	.013	-	-.881 <sup>a</sup>
<i>me</i> <sub>-19</sub>	-	.530	-	-.009	-	-.847 <sup>a</sup>
$\bar{R}^2$	.483	.562	.491	.638	.594	.745
<i>D.W</i>	2.04	2.22	1.91	2.05	1.98	2.33
<i>d.f.</i>	16.40	16.27	16.40	16.27	16.40	16.27

<sup>a</sup>Significant at the 99 percent level.<sup>b</sup>Significant at the 95 percent level.<sup>c</sup>Significant at the 90 percent level.

The results presented here are not without problems. The procedure used to generate expectations is still far from perfect. There remains a potential aggregation problem with different economic agents potentially having systematically different expectations. I am attempting to approximate an aggregate un-

observable variable. In addition, expectations are continually revised and updated based on new information, while in this study I can only attempt to model expectations at a particular point in time. Finally, economic agents may also have more information than it has been assumed they have. For example,

expectations are also undoubtedly based in part on variables not considered here, such as elections, wars, and rumors.

## REFERENCES

- Attfield, C. L. F., Demery, D. and Duck, N. W., "Unanticipated Monetary Growth, Output and the Price Level: U.K. 1946-77," *European Economic Review*, December 1981, 16, 367-85.
- Barro, Robert, J., "Unanticipated Money Growth and Unemployment in the United States," *American Economic Review*, March 1977, 67, 101-15.
- \_\_\_\_\_, "Unanticipated Money, Output, and the Price Level in the United States," *Journal of Political Economy*, August 1978, 86, 549-80.
- \_\_\_\_\_, "A Capital Market in an Equilibrium Business Cycle Model," *Econometrica*, November 1980, 48, 1393-417.
- \_\_\_\_\_, and Hercowitz, Z., "Money Stock Revisions and Unanticipated Money Growth," *Journal of Monetary Economics*, April 1980, 6, 257-68.
- \_\_\_\_\_, and Rush, Mark, "Unanticipated Money and Economic Activity," in S. Fischer, ed., *Rational Expectations and Economic Policy*, Chicago: University of Chicago Press, 1980, 23-48.
- Begg, David K. H., *The Rational Expectations Revolution in Macroeconomics*, Baltimore: Johns Hopkins University Press, 1982.
- Grossman, Jacob, "Nominal Demand Policy and Short-Run Fluctuations in Unemployment and Prices in the United States," *Journal of Political Economy*, October 1979, 87, 1063-85.
- King, Robert G., "Monetary Information and Monetary Neutrality," *Journal of Monetary Economics*, March 1981, 7, 195-206.
- Leiderman, Leonard, "Macroeconomics Testing of the Rational Expectations and Structural Neutrality Hypotheses for the United States," *Journal of Monetary Economics*, January 1980, 6, 69-82.
- Ljung, C. M. and Box, G. E. P., "Measure of Lack of Fit in Time Series Models," *Biometrika*, 1978, 65, 297-303.
- Lucas, Robert, E. Jr., "Expectations and the Neutrality of Money," *Journal of Economic Theory*, April 1972, 4, 103-24.
- McCallum, Bennet T., "Rational Expectations and Macroeconomic Stabilization Policy," *Journal of Money, Credit, and Banking*, November 1980, 12, 716-46.
- Makin, John H., "Anticipated Money, Inflation Uncertainty and Real Economic Activity," *Review of Economics and Statistics*, February 1982, 64, 126-34.
- Mishkin, Frederic S., (1982a) "Does Anticipated Monetary Policy Matter? An Econometric Investigation," *Journal of Political Economy*, February 1982, 90, 22-51.
- \_\_\_\_\_, (1982b) "Does Anticipated Aggregate Demand Matter? Further Econometric Results," *American Economic Review*, September 1982, 72, 788-802.
- Sargent, Thomas J., and Wallace, Neil, "Rational Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule," *Journal of Political Economy*, April 1975, 83, 241-54.
- Sheffrin, Steven M., "Unanticipated Money Growth and Output Fluctuations," *Economic Inquiry*, January 1979, 17, 1-13.
- Small, David H., "Unanticipated Money Growth and Unemployment in the United States: Comment," *American Economic Review*, December 1979, 69, 996-1003.
- Weiss, Laurence, "The Roles for Active Monetary Policy in a Rational Expectation Model," *Journal of Political Economy*, April 1980, 88, 221-33.

# Inflation and Reputation

By DAVID BACKUS AND JOHN DRIFFILL\*

Economists have speculated for years about the source of the apparent inflationary bias of market economies. In the 1960's the Phillips curve supplied part of an explanation: with a stable tradeoff between inflation and output, governments might reasonably choose a positive rate of inflation, even if they find inflation distasteful, in order to raise output.

Natural rate theories changed all that. If the government can only raise output with surprise inflation, then systematic expansionary policy will generate inflation but fail to raise output. If a stable price level is desirable, the only sensible policy is zero inflation. The question then is why government policy has tolerated persistent high rates of inflation over the past decade or so.

One answer is that zero inflation is not a credible policy if the government is known to care about output. This has arisen as "dynamic inconsistency" in Finn Kydland and Edward Prescott (1977) and as inferiority of Nash solutions in Robert Barro and David Gordon (1983a,b), but generally reflects the fact that noncooperative equilibria need not be Pareto optimal.

Another line of argument runs that governments, exploiting the short memories of voters, overheat the economy prior to elections. (William Nordhaus, 1975, provides such a model of "political business cycles" and Gerald Kramer, 1971, and Ray Fair, 1978, report evidence that voters seem to be shortsighted; Henry Chappell, 1983, gives a conflicting view.) This mechanism, however, only leads to an inflationary bias if the Phillips curve is nonlinear, so that high output

raises inflation by more than low output reduces it.

In the following sections we extend the work of Barro and Gordon to a situation in which the public is uncertain about the preferences of the government: in particular, whether it cares about unemployment and output. Thus, when the government announces its intention to fight inflation regardless of the output cost, the public is uncertain whether this is in fact the case, or whether it is simply an attempt to manipulate their expectations. This analysis of reputation, based on David Kreps and Robert Wilson (1982b), provides a useful formalization of the credibility problem faced by macroeconomic policymakers, and stressed repeatedly by William Fellner (1982).

One feature of the model is that tight macroeconomic policy aimed at eliminating inflation will reduce output below the natural rate if the public thinks the government may inflate. Peter Howitt (1982) makes a similar point without elaborating on the source of the public's skepticism. The critical element in our model is the public's lack of information about the government: even if the government is serious about combating inflation, the public cannot know this with certainty. Completely credible noninflationary policy is generally not possible.

A second feature of the analysis is that government policy is dynamically consistent: in equilibrium the government always finds it optimal to stick to its initial plan. By treating policy as a dynamic game and applying Kreps and Wilson's (1982a) notion of sequential equilibrium, we avoid the "inconsistency of optimal plans" that has plagued studies that view policy as an optimization problem.

At the same time we explain the political business cycle without recourse to voter naivete. Even governments that care about employment will tend, at the start of their terms of office, to act as if they do not in order to keep alive in the mind of the public

\*Departments of Economics, Queen's University, Kingston, K7L 3N6 Canada, and University of Southampton, Southampton, S09 5NH England, respectively. Backus thanks the SSHRC of Canada for financial support.

the possibility that they will fight inflation at all costs. Of course, such governments will always inflate near the end of their terms in an attempt to raise output. What is more, it will work: output rises (probabilistically) as long as the government still has a reputation for toughness. The public acts rationally throughout: it simply does not know what the government plans to do.

These ideas are developed in the rest of the paper. Section I reviews the Barro-Gordon model. The analysis of Kreps and Wilson is applied to this model in Section II. A numerical example and discussion follow. The final section contains general remarks about the strengths and limitations of the analysis.

### I. The Barro-Gordon Analysis

Barro and Gordon (1983b) characterize macroeconomic policy as a game. Output is determined by a Phillips curve with the natural rate property:

$$y = y_n + (x - x^e),$$

where  $y$  is output,  $y_n$  is the natural rate, and  $x$  and  $x^e$  are actual and expected inflation. The government likes output and dislikes inflation, which we may formalize with the one-period payoff function

$$\begin{aligned} (1) \quad u_g(x, x^e) &= -\frac{1}{2}ax^2 + b(y - y_n) \\ &= -\frac{1}{2}ax^2 + b(x - x^e). \end{aligned}$$

The public, on the other hand, resists being fooled; that is, they maximize

$$(2) \quad u_p(x, x^e) = -(x - x^e)^2.$$

As Barro and Gordon (1983a) argue, these payoffs are consistent with the government and the public having identical preferences. If the natural rate of unemployment is too high, perhaps because of taxes or externalities, then everyone might agree that maximizing (1) is desirable. But for individual agents, aggregate inflation and output are givens; the best they can do is forecast inflation

accurately. The game consists then of the government choosing  $x$  and the public choosing  $x^e$ , with payoffs given by equations (1) and (2).

Now consider, as do Barro and Gordon, the Nash solution to the game in which both players move simultaneously. With both government and public maximizing given the other's decision, the solution is

$$x = b/a \quad \text{and} \quad x^e = b/a.$$

The model explains inflation as the Nash equilibrium to a policy game. The payoffs are  $u_g = -(1/2)b^2/a$  and  $u_p = 0$ , which is Pareto inferior to the zero-inflation solution ( $x = x^e = 0$ ) in which  $u_g = u_p = 0$ .

Barro and Gordon argue persuasively that the inefficiency stems from the government's inability to commit itself to a noninflationary policy. Suppose, for example, that the government were able to move first, committing itself to a particular rate of inflation. Then an intelligent government would choose  $x$  taking into account the public's response (namely,  $x^e = x$ ) and pick zero inflation.

But when the government cannot make prior commitments, it faces a problem in convincing the public that it will, in fact, choose zero inflation. For if the public believes this, the government has an incentive to inflate at rate  $x = b/a$ , thereby raising output and the government's payoff. Using the normalization  $a = b = 2$ , we can represent policy as a matrix game with two strategies ( $x, x^e = 0$  or 1), and payoffs

		public	
		$x^e = 0$	$x^e = 1$
government	$x = 0$	0, 0	-2, -1
	$x = 1$	1, -1	-1, 0.

(In each ordered pair the government's payoffs are listed first.) The problem from this point of view is that  $x = 1$  is a dominant strategy for the government: the payoffs are larger regardless of what the public does. The public, therefore, sensibly expects the government to inflate. The result is the Pareto-inferior solution  $x = x^e = 1$ .

Barro and Gordon then consider the possibility that the government can establish a reputation for avoiding inflation if the game is repeated infinitely many times. Let the government's payoff for the repeated game be

$$J_g = \sum_{t=0}^{\infty} c^t u_g(x_t, x_t^e), \quad 0 < c < 1,$$

with  $c$  a discount factor. Then Barro and Gordon (1983b), applying results similar to James Friedman (1971; 1977, ch. 8), show that the strategies

$$x_t^e = \begin{cases} x^* & \text{if } x_{t-1} = x^* \\ 1 & \text{otherwise} \end{cases}$$

$$x_t = x^*$$

do *not* constitute a Nash equilibrium when  $x^* = 0$ . The problem is that the government's benefit from cheating (the extra payoff of +1 when  $x = 1$  and  $x^e = 0$ ) is equal to its cost (the loss of 1 when  $x^e = 1$  and the government returns to zero inflation). But since the latter comes later, the discount factor ensures that cheating is a superior strategy and  $x^* = 0$  is therefore not a Nash equilibrium with the given "punishment" strategy. They go on to argue that some positive rates of inflation  $x^*$  can be sustained as Nash equilibria, and it seems clear that even zero could be sustained with a longer punishment interval if  $c$  is not too small.

A weakness of this analysis is that the punishment strategy played by the private sector (punish the government by playing  $x_t^e$  equal to 1 if  $x_{t-1}$  is not zero) is essentially arbitrary. Further, the equilibrium which is sustained depends critically on the form this punishment strategy takes. The infinite horizon game has multiple Nash equilibria, with no mechanism for choosing among them. The Kreps-Wilson (1982b) analysis of reputation, to which we now turn, avoids these problems and illuminates a number of other issues as well.

## II. Reputational Equilibrium

Consider now the possibility that the government may behave in one of two ways: it may behave as if it is rationally attempting to maximize the utility function (1) (a "wet" government); or it may behave as if it is committed irrevocably to pursuing a zero-inflation policy (a "hard-nosed" (H-N) government). Wet governments therefore have payoffs

		public	
		$x^e = 0$	$x^e = 1$
wet government	$x = 0$	0	-2
	$x = 1$	1	-1,

as in the previous section. Hard-nosed governments, however, behave as if their payoffs are

		public	
		$x^e = 0$	$x^e = 1$
H-N government	$x = 0$	0	0
	$x = 1$	-1	-1,

which we might derive by setting  $b = 0$  in the government payoff function (1). They therefore have no incentive to inflate, regardless of expectations. The public's payoffs are given by equation (2):

		public	
		$x^e = 0$	$x^e = 1$
government	$x = 0$	0	-1
	$x = 1$	-1	0.

The crux of the analysis is that the public does not know which type of government behavior it faces. As a result, even a wet government may choose not to inflate. By resisting inflation it develops a reputation for being hard nosed which it hopes will discourage expectations of inflation in the future. In this section we examine such a reputational equilibrium in a finitely repeated version of the Barro-Gordon policy game when the public is uncertain about the government's behavior. The analysis is identical to Kreps and Wilson (1982b, Sec. 3) in all essen-

tial respects. The solution concept is Kreps and Wilson's (1982a) sequential equilibrium, which enables us to find the solution recursively, starting with the final period.

The central feature of the model is the government's ability to manipulate its reputation. The government enters period  $t$ , say, with a reputation  $p_t$  equal to the public's probability that the government is hard nosed. By assumption both the government and the public know  $p_t$ . Both players then choose their best strategies, given the other's strategy and the impact of current behavior on the next period's reputation. The probability  $p_t$  is then revised in light of observed behavior according to Bayes' rule.

Each player's strategy is usefully characterized as a probability of playing zero in a mixed strategy: denoted  $z_t$  for the public and  $y_t$  for the government. Then the government's reputation next period,  $p_{t+1}$ , is zero if it inflates this period (or has ever inflated in the past) since  $H-N$  governments never inflate. Given no inflation, Bayes' rule gives the probability as

$$\begin{aligned} p_{t+1} &= \text{prob}(H-N|x_t=0), \\ &= \text{prob}(H-N \text{ and } x_t=0)/\text{prob}(x_t=0), \\ &= \text{prob}(x_t=0|H-N)\text{prob}(H-N) \\ &\quad / [\text{prob}(x_t=0|H-N)\text{prob}(H-N) \\ &\quad + \text{prob}(x_t=0|\text{wet})\text{prob}(\text{wet})], \end{aligned}$$

or

$$(3) \quad p_{t+1} = p_t / [p_t + (1-p_t)y_t].$$

In this game, as in the version of the chain store paradox analyzed by Kreps and Wilson (1982b, Sec. 2), the probability  $p_t$  is a sufficient statistic for past play and contains all the relevant information needed by the players to make optimal decisions.

Consider now the solution of the game. In the final period,  $T$ , a  $H-N$  government will always play  $x_T=0$ . The expected return for

a wet government is

$$\begin{aligned} (4) \quad J_g(T, p_T) &= z_T[y_T(0) + (1-y_T)(1)] \\ &\quad + (1-z_T)[y_T(-2) + (1-y_T)(-1)] \\ &= (2z_T-1)-y_T. \end{aligned}$$

Since this is declining in  $y_T$ , a wet government will always inflate in the last period:  $y_T=0$ . Similarly, the public's expected payoff is

$$\begin{aligned} J_p(T, p_T) &= z_T[p_T(0) + (1-p_T)(-1)] \\ &\quad + (1-z_T)[p_T(-1) + (1-p_T)(0)] \\ &= z_T(2p_T-1) - p_T. \end{aligned}$$

Thus if  $p_T > 1/2$  the public plays  $z_T=1$  ( $x_T^e=0$ ), if  $p_T < 1/2$  it plays  $z_T=0$ , and if  $p_T=1/2$  the public is completely indifferent about  $z_T$ . The equilibrium strategy in this case ( $z_T=1/2$  when  $p_T=1/2$ ) will be derived below from the equilibrium conditions for the preceding period. The value to the government to playing the game in the last period is therefore

$$v_g(T, p_T) = \begin{cases} 1 & \text{if } p_T > 1/2, \\ 0 & \text{if } p_T = 1/2, \\ -1 & \text{if } p_T < 1/2. \end{cases}$$

The value to the public is

$$v_p(T, p_T) = \max(-p_T, p_T-1).$$

In period  $T-1$ , the government must consider the impact of its behavior on its reputation in the final period. The expected two-period payoff is

$$\begin{aligned} J_g(T-1, p_{T-1}) &= z_{T-1}[y_{T-1}(0) + (1-y_{T-1})(1)] \\ &\quad + (1-z_{T-1})[y_{T-1}(-2) + (1-y_{T-1})(-1)] \\ &\quad + y_{T-1}v_g(T, p_T) + (1-y_{T-1})(-1). \end{aligned}$$

The last term reflects the fact that if the government inflates in period  $T-1$ , which it does with probability  $(1 - y_{T-1})$ , then its reputation is blown; the public will expect inflation in the final period and the government's payoff in that period is  $-1$ . The penultimate term is the probability of playing zero actual inflation in  $T-1$ , and then collecting the payoff in  $T$  associated with a reputation  $p_T$ , where  $p_T$  is given by equation (3). The expression reduces to

$$(5) \quad J_g(T-1, p_{T-1}) = 2z_{T-1} - 2 \\ + y_{T-1}v_g(T, p_T).$$

The public's two-period payoff is

$$J_p(T-1, p_{T-1}) = z_{T-1}(2q_{T-1} - 1) - q_{T-1} \\ + q_{T-1}v_p(T, p_T),$$

where  $q_{T-1} = p_{T-1} + (1 - p_{T-1})y_{T-1}$ .

The government now chooses  $y_{T-1}$  to maximize (5) subject to (3). For  $p_{T-1} > 1/2$ , this implies  $y_{T-1} = 1$ , hence  $p_T = p_{T-1} > 1/2$ ,  $v_g(T, p_T) = 1$ , and  $z_T = 1$ . That is, the government plays  $x_{T-1} = 0$  with certainty; its reputation does not change, but it is already sufficient to ensure that the public does not expect inflation in the last period. For  $0 < p_{T-1} < 1/2$ , the government plays zero inflation with probability

$$y_{T-1} = p_{T-1}/[1 - p_{T-1}].$$

If by chance it fails to inflate, its reputation for being hard nosed rises in the next period to  $1/2$ . Since  $p_{T-1} < 1/2$  it is clear that  $y_{T-1}$  is strictly less than one. But since  $y_{T-1}$  maximizes (5), this can only be true if  $v_g(T, p_T) = 0$ . From (4) we see then that  $z_T$  must be  $1/2$ , as we claimed earlier.

If  $p_{T-1}$  is exactly  $1/2$  then  $y_{T-1}$  is one and it appears that any value of  $z_T$  between one-half and one is consistent with equilibrium. We assume in this case that  $z_T = 1/2$ . This assumption is analogous to one made by Kreps and Wilson and has no material effect on the results.

The solution of the game as described has the property that both  $(x_T, y_T)$  and  $(x_{T-1}, y_{T-1}, x_T, y_T)$  are Nash equilibria. This recursive structure, which Kreps and Wilson (1982a) have labeled "sequential equilibrium," imposes a condition on the solution analogous to the principle of optimality. In period  $T-1$ , we only consider period  $T$  strategies which are themselves Nash equilibria. As a result, the equilibrium is dynamically consistent by construction.

With similar reasoning the solution can be extended to earlier periods. Equilibrium behavior is conveniently summarized as follows. (i) In period  $t$ , the private sector plays zero expected inflation with probability  $z_t$  given by

$$z_t = \begin{cases} 1 & \text{if } p_t > (1/2)^{T-t+1}, \\ 1/2 & \text{if } p_t = (1/2)^{T-t+1}, \\ 0 & \text{if } p_t < (1/2)^{T-t+1}. \end{cases}$$

(ii) A wet government plays actual inflation equal to zero with probability  $y_t$  given by

$$y_t = \begin{cases} 1 & \text{if } p_t > (1/2)^{T-t}, \\ \frac{p_t}{1-p_t} \frac{1 - (1/2)^{T-t}}{(1/2)^{T-t}} & \text{if } 0 < p_t \leq (1/2)^{T-t}, \\ 0 & \text{if } p_t = 0. \end{cases}$$

(iii) The probability of the government being hard nosed is revised in accordance with Bayes' rule:

$$p_{t+1} = \begin{cases} \frac{p_t}{p_t + (1-p_t)y_t} & \text{if } x_t = 0, \\ 0 & \text{if } x_t = 1 \\ & \text{or } p_t = 0. \end{cases}$$

(iv) The expected payoff to a wet government on entering stage  $t$  of the game with reputation  $p_t$  is given by

$$J_g(t, p_t) = 2z_t - (T - t + 1) \\ + y_t[v_g(t+1, p_{t+1}) + (T - t - 1)],$$



where  $v_g(t+1, p_{t+1})$  is the value of the game next period conditional on not inflating in period  $t$ . In equilibrium the value function for a wet government is therefore

$$v_g(t, p_t) = \begin{cases} t-T-1 & \text{if } 0 < p_t < (1/2)^{T-t+1}, \\ t-T & \text{if } p_t = (1/2)^{T-t+1}, \\ t-T+1 & \text{if } (1/2)^{T-t+1} < p_t \leq (1/2)^{T-t}, \\ t-T+2+i & \text{if } (1/2)^{T-t-i} < p_t \leq (1/2)^{T-t-i-1}, \end{cases}$$

for  $i = 0, 1, \dots, T-t-1$ ,  
and  $t = 1, 2, \dots, T-1$ .

### III. Reputation and Dynamically Consistent Policy: An Example

To get an idea as to what kinds of behavior are implied by the theory, let us look at an example. Suppose a wet government comes to power with a five-year term ( $T=5$ , with no possibility of reelection) and that at the beginning of its term it is strongly suspected of being wet. To be specific, let us say that it is believed to be hard nosed with probability lying somewhere between  $1/16$  and  $1/32$ , although none of the qualitative conclusions depend on these values.

The play progresses as follows. In period 1, the public chooses  $x_1^e = 0$  with certainty and the government chooses  $x_1 = 0$  with probability  $15p_1/(1-p_1) < 1$ . If  $x_1 = 0$  is actually played, the game continues with the government's reputation enhanced ( $p_2 = 1/2^4 = 1/16$ ). If the government inflates, its reputation is ruined, and the equilibrium is  $x = x^e = 1$  for the rest of the game.

In later periods, if the government has not yet inflated, its reputation rises just enough to induce the public to choose zero expected inflation with probability  $1/2$ . Three points are noteworthy. (i) Reputation is only enhanced ( $p_t > p_{t-1}$  given that  $x_{t-1} = 0$ ) if the government plays  $x_{t-1} = 0$  with probability less than one. Acquiring a reputation thus involves taking a risk. (ii) In each period after the first, given that actual inflation was zero in the preceding period, the public

randomizes with a constant probability of  $1/2$ . Thus there is a positive probability of getting  $x^e = 1$ ,  $x = 0$  and therefore a recession. The revised estimates that the government is hard nosed are not enough to discourage the public completely from playing  $x^e = 1$ . (iii) The probability that a wet government survives until the last period without ruining its reputation is just equal to  $p_1/(1-p_1)$ , so it pays to have a good reputation.

Let us consider now the problem of dynamic inconsistency and the definition of optimal policy. As Kydland and Prescott showed, the *ex ante* optimal policy is typically dynamically inconsistent, and therefore not credible. But the outcome of the best consistent policy is frequently worse than the *ex ante* optimal policy if the latter is credible.

Our own solution is, by construction, dynamically consistent and credibility is conveniently summarized by the reputation,  $p$ . It is easy to see that the sequential equilibrium is the best credible policy. In our example, the payoff to the government of following the consistent sequential equilibrium policy is  $v_g(1, p_1) = -3$ . Alternatively, suppose the government played  $x = 1$  in every period. The public, given  $1/16 < p_1 < 1/32$ , would play  $x_1^e = 0$  and  $x_t^e = 1$  ( $t = 2, \dots, 5$ ), giving the government a payoff of  $-4$ . The outcome involves actual inflation in every period, which is a surprise only in period 1. (If the private sector anticipated  $x_1 = 1$ , the government payoff would be  $-5$ .) Finally, if the government announced that it would never inflate, the outcome depends on whether it is believed or not. If  $p_1$  truly captures the public's beliefs, then the outcome is zero actual inflation in each period, but zero expected inflation with certainty in period 1 and with probability  $1/2$  thereafter, giving the government an expected payoff of  $-4$ . If the announced zero-inflation policy were believed (because of an associated constitutional amendment, for example), then the government's payoff would be zero.

It is clear, then, that the sequential equilibrium is dominated only by the fully believed zero-inflation commitment, given the behavior of the public and their beliefs about

the government. It is always at least as good as pursuing a zero-inflation policy "come what may," in the face of a poor reputation. Thus, in the presence of public skepticism ( $p_1 < 1$ ), and in the absence of irrevocable commitments, the sequential equilibrium is at least as good as (and usually better than) the time-inconsistent *ex ante* optimal policy of setting  $x_t = 0$  in all periods. The concept of sequential equilibrium removes the ambiguity from the definition of optimal policy.

The analysis also sheds light on the results of Barro and Gordon. If a government is optimizing over a long time horizon ( $T$ ), and its initial reputation is "good" in the sense that  $p_1$  is much larger than  $(1/2)^T$ , then the solution will have the following property. There will be an initial period in which zero inflation is expected, and zero actual inflation occurs. The first period in which there is a departure from this pattern is period  $n$ , where  $n$  is the smallest integer for which  $(1/2)^{T-n} > p_1 > (1/2)^{T-n+1}$ . In this period the government will begin to create inflation with some nonzero probability. In subsequent periods, conditional on having observed zero inflation, the private sector will expect zero inflation with probability  $1/2$ .

As  $T$  increases for given  $p_1$ ,  $n$  increases also. In contrast to Barro and Gordon, in a game with a sufficiently long time horizon this analysis leads to the conclusion that for any nonzero initial reputation ( $p_1 > 0$ ) there will be an initial period (which tends to infinity as the horizon tends to infinity) in which zero inflation is the equilibrium outcome. This is supported by the "punishment strategy" for the private sector which makes  $x_t^e = 1$  for all  $t > s$  if  $x_s = 1$ . This strategy is rational if the government's behavior is used to draw inferences about its preferences. By contrast, Barro and Gordon (1983b) assume that the public "punishes" a deviant government for one period, without rationalizing this assumption, and argue that zero inflation cannot be supported in an infinite-horizon game except by the use of rules to lend credibility to policy announcements. Our analysis would change somewhat if we introduced a discount factor. As in Barro and Gordon, a small discount factor makes it harder to sustain low inflation.

#### IV. Final Remarks

The Kreps-Wilson analysis fits many of the observed features of macroeconomic policy quite well. First, it is commonplace to hear politicians reassure us that they are serious about beating inflation. These statements are correctly regarded with skepticism, since both hard-nosed and wet governments have an incentive to establish reputations for being tough—that is, raise  $p$ . Conversely, governments frequently complain that their actions are thwarted by the "mistaken" expectations of labor unions, big business, and the gnomes of Zurich. Note, for example, that even a hard-nosed government will suffer persistent output losses half the time as the public randomizes, if its initial reputation is bad ( $p$  is small). Wet governments will also induce recessions until, by chance, they reveal themselves to be wet. From then on the inflationary equilibrium ( $x = x^e = 1$ ) results.

Second, the model provides an account of the political business cycle without relying on voter myopia. At the same time it explains the inflationary bias of policy without recourse to nonlinearities in the Phillips curve. In the Nordhaus model, governments deflate early in their terms and inflate later since voters place more weight on events immediately preceding the next election. The strategy successfully raises output because the public is doubly myopic: they forget the low output early in the term and they fail to predict the inflation later. In the Kreps-Wilson framework, inflation at the end of the term is the rational response of a government that cares about employment. It works, on average, because the public is uncertain about the government's true character. Voters are not myopic; they simply do not have all the information.

The logic is just the opposite of conventional theory of the political business cycle. Instead of having a government create a pre-election boom in order to increase its chance of reelection, our analysis generates a pre-election boom as the solution to a game with a wet lame-duck government. In fact, if there were a chance of reelection, the incentive to preserve reputation may actually restrain the spending spree.

Third, the analysis suggests that governments may try to appoint central bankers with reputations for fighting inflation, even if their own preferences place positive weight on employment. By doing so they minimize the costs associated with uncertainty about policy ( $v_p$  is highest when  $p$  is zero or one) and with the credibility problem wet governments have with noninflationary policies. Autonomous central banks thus act as a precommitment device which may help to make noninflationary policies more credible and less costly.

Despite these apparent strengths of the analysis, a few caveats are in order. On the technical side, the assumptions that there are only two choices for inflation may be undesirably restrictive. The lack of dynamics relating inflation and output to their past values is also troublesome. The possibility of an intransigent public sector is discussed in our earlier paper (1984).

We also have some doubt that the model explains why we have had relatively high inflation during the past fifteen years, but not before. James Tobin, for one, disagrees that inflation was a policy choice derived from a desire to raise employment.

Today a widespread version of recent history is that governments deliberately sought higher inflation in order to reduce unemployment....As an explanation of recent inflation in the United States this account is enormously exaggerated....The 1966-69 ride up the Phillips curve was not a conscious choice of novel macroeconomic strategy but a timeworn political decision about wartime finance. Against the advice of his Keynesian advisors, President Johnson chose for his own reasons of domestic and international politics not to ask Congress for increased taxes to finance his ill-starred escalation of the conflict in Indochina.

[1981, pp. 21-22]

But whatever the origin of the inflation, we think the model helps to explain why disinflation took so long and was so painful. By the mid-1970's, the public was highly skeptical of each new attempt to fight inflation,

since so many attempts had been abandoned in the past. Presumably even tough-minded policymakers faced a doubtful public. As a result, governments who cared about employment were often forced (probabilistically) to continue inflationary policies. Governments who wished only to stop inflation could not easily persuade the public of this fact, and therefore induced severe protracted recessions when they tried.

## REFERENCES

- Backus, David and Driffill, John, "Rational Expectations and Policy Credibility Following a Change in Regime," Discussion Paper, Queen's University, June 1984.
- Barro, Robert and Gordon, David, (1983a) "A Positive Theory of Monetary Policy in a Natural-Rate Model," *Journal of Political Economy*, August 1983, 91, 589-610.
- \_\_\_\_\_, and \_\_\_\_\_, (1983b) "Rules, Discretion, and Reputation in a Model of Monetary Policy," *Journal of Monetary Economics*, July 1983, 12, 101-21.
- Chappell, Henry, "Presidential Popularity and Macroeconomic Performance: Are Voters Really so Naive?," *Review of Economics and Statistics*, August 1983, 65, 385-92.
- Fair, Ray, "The Effect of Economic Events on Votes for President," *Review of Economics and Statistics*, May 1978, 60, 159-73.
- Fellner, William, "Towards a Reconstruction of Macroeconomics," in Martin Baily and Arthur Okun, eds., *The Battle Against Unemployment and Inflation*, 3d ed., New York: W. W. Norton, 1982.
- Friedman, James, "A Noncooperative Equilibrium for Supergames," *Review of Economic Studies*, January 1971, 28, 1-12.
- \_\_\_\_\_, *Oligopoly and the Theory of Games*, Amsterdam: North-Holland, 1977.
- Howitt, Peter, "Anti-Inflation Policy with a Skeptical Public: A Comment on the Meyer-Webster Paper," *Carnegie-Rochester Conference Series on Public Policy: Economic Policy in a World of Change*, Autumn 1982, 17, 109-14.
- Kramer, Gerald, "Short-term Fluctuations in U.S. Voting Behavior, 1896-1964," *American Political Science Review*, March 1971, 65, 131-43.

# The Competitive Effects of Vertical Agreements?

By WILLIAM S. COMANOR AND H. E. FRECH III\*

For many years, there were few distinctions drawn between horizontal and vertical agreements. Both were considered anticompetitive and subject to *per se* condemnation under the antitrust laws. Recently, however, this approach has come under attack, and what was once the conventional wisdom is no longer so. Indeed, there is growing acceptance of the view that vertical agreements can rarely have anticompetitive consequences. *Per se* legality would then be the appropriate standard.

In this paper, we investigate the competitive implications of a particular vertical agreement: the imposition of exclusive dealing requirements by a manufacturer on his distributors. However, to maintain the focus of the analysis, we do not consider ultimate welfare gains or losses.

In an early application of economic analysis to this practice, Aaron Director and Edward Levi (1956) suggest that exclusive dealing would be anticompetitive if it raised entry costs for rivals. Our object, following this conjecture (see their p. 293), is to examine the market conditions under which exclusive dealing impedes entry.

Howard Marvel (1982) dealt with the practice of exclusive dealing. He provides an efficiency rationale for exclusive dealing, ignores the prospect that anticompetitive effects may follow, and concludes that "exclusive dealing ought therefore to be treated as legal, *per se*" (p. 25). This paper examines the possible anticompetitive effects neglected by Marvel.

## I. Market Conditions for Exclusive Dealing

At this point, we construct a simple model of exclusive dealing arrangements. Our object is not a search for generality, but rather to illustrate both the conditions and the manner by which exclusive dealing may have anticompetitive effects.

Our first assumption is that there exists a single dominant manufacturer producing a single product at constant unit costs  $c$  in the relevant range of output. There are two classes of consumers for the firm's product. Members of the first group, indicated as Class A, believe it superior to any rival brand. These consumers are identical and purchase the dominant firm's product rather than a competing brand so long as its price to them is no higher than  $\alpha$  above the rival's price. Their individual demand curves for this product therefore lie above the corresponding demand curves for all other products by the same amount  $\alpha$ , which reflects the per unit value of Class A consumers' brand preferences.

The remaining consumers, members of Class B, view the products of all sellers as identical. For convenience, we assume that the individual demand curve of Class B consumers is identical to that of Class A consumers for all brands except that of the dominant manufacturer. The aggregate demand curve therefore depends on the number of consumers in each class as well as the size of  $\alpha$ .

Consumers obtain the product through independent distributors. Since resale between customers is possible, discrimination between the two classes of consumers is not feasible, and the same price is charged to all.

We assume further that imperfect competition exists among established distributors. Scale economies in distribution create segmented local markets. As a result, established distributors do not compete on a large,

\*Professors of Economics, University of California, Santa Barbara, CA 93106. We thank the Committee on Research of the Academic Senate, University of California-Santa Barbara, for financial support. Valuable comments were received from seminars at the universities of Chicago, Northwestern, Toronto, and Illinois, the Federal Trade Commission, and at our own university.

dense plain. They have some degree of market power and set distribution margins above their marginal costs.

Distribution margins in these circumstances are determined by spatial competition of the type modelled by Steven Salop and others.<sup>1</sup> For our analysis, a change in either the manufacturer's price or his policy regarding exclusive or nonexclusive dealing would affect both demand and costs for the distribution function. However, it would be extremely complex and lead us away from our main subject to incorporate formally these effects. We therefore adopt the simplifying assumption that the number of distributors and the imperfectly competitive distribution margin is constant with respect both to the manufacturer's price and policy regarding exclusive dealing. This margin is indicated by  $\gamma$ .

This simplifying assumption actually comports with reality better than one might expect. Due to the presence of economies of scope, distributors handle many, perhaps thousands, of products. Therefore, the actual number of distributors is unlikely to be much affected by what happens with any single product. And as a result, equilibrium margins which depend strongly on the number and location of rivals would be relatively stable.

Distribution is a multiproduct activity that is subject to economies of scope.<sup>2</sup> These economies imply that

$$(1) \quad C(q_1, \dots, q_n) < C(q_1) + C(q_2) + \dots + C(q_n),$$

where  $C(q_1, \dots, q_n)$  indicates the total costs of distributing a vector of  $q_i$  goods. The terms on the right-hand side of expression (1) indicate the total costs of distributing the same vector of goods separately.  $C(q_1)/q_1$  is therefore the unit cost of distributing the first good without realizing economies of scope. It represents the minimum distribution margin

that can be charged by single-product distributors, and is indicated by  $\delta$ . In the presence of economies of scope,  $\delta$  exceeds  $\gamma$ .

## II. Distribution without Exclusive Dealing

First, consider the case where the manufacturer ignores the prospect of entry. Since established distributors set a distribution margin of  $\gamma$ , the market demand curve facing the manufacturer is the consumer demand curve shifted downward by that amount. The dominant manufacturer maximizes profits, given that demand curve.

Now consider the case where a dominant manufacturer takes the prospect of entry into account. To focus the discussion, let the entrant have the same constant manufacturing costs as the established firm. Although Class B consumers view his product as identical to that of the original manufacturer, Class A consumers consider his product less valuable so their aggregate demand price is lower by  $\alpha$  per unit.

While there are surely strategic interactions between entrant and established firm, we abstract from these issues and assume that the entrant behaves competitively regardless of the actions of the dominant firm.<sup>3</sup> He enters if he can obtain a price of  $c$  for his product and sets this price in all circumstances. Not only does this assumption simplify the analysis of oligopolistic interdependence between entrant and established firm, it represents that strategic choice on the part of the entrant that is most conducive to competition.

Even in these circumstances, there are two possible strategies which can be adopted by the original manufacturer. The first, a low-price strategy, is to set his price at  $c$ . Entry is prevented and the entire market retained by the original manufacturer. This strategy, however, leads to zero profits. Since a better one exists, a low-price strategy is not adopted.

Following a high-price strategy, the firm sets a higher limit-entry price that retains sales to Class A consumers but sacrifices

<sup>1</sup>Salop (1979a). See also Jack Hirshleifer (1980, pp. 363-76).

<sup>2</sup>For a discussion of economies of scope, see Robert Willig (1979, pp. 346-47).

<sup>3</sup>See F. M. Scherer (1980, pp. 229-66).

sales of Class B consumers to the entrant. The manufacturer's price is now  $(c + \alpha)$ , which leads to a final price to consumers of  $(c + \alpha + \gamma)$ . The entrant's price is merely  $c$  and the resale price of his product is  $(c + \gamma)$ . These prices are summarized below:

(2)	$P_M = c + \alpha$	Manufacturer's price
	$P'_M = c + \alpha + \gamma$	Distributor's price of original manufacturer's product
	$P_E = c$	Entrant's price
	$P'_E = c + \gamma$	Distributor's price of the entrant's product.

Within the confines of this model, sales to Class A consumers are of the original manufacturer's product while sales to Class B consumers are of the entrant's product. The original manufacturer's profits result from his product differentiation advantages, represented by  $\alpha$ , while the distributor's profits, if any, rest on his locational or other advantages, indicated by  $\gamma$  less distribution costs.

### III. Distribution with Exclusive Dealing

Even with exclusive dealing, contracts remain simple. Complex contracts that effectively integrate the manufacturer and distributor are assumed too costly to reach.

After the dominant manufacturer has imposed exclusive dealing requirements, the new entrant must distribute his product through an alternate channel and bear higher distribution costs. The relevant prices for the entrant's product are now

(3)	$P_E = c$	Entrant's price
	$P'_E = c + \delta$	Resale price of the entrant's product.

The original manufacturer can benefit from the differential costs of distribution. This factor is the key to understanding the possible anticompetitive effects of exclusive dealing.

Again, two strategies are possible. Following a low-price strategy, the manufacturer sells to both classes of customers. In this case, his final price must be no higher than the corresponding price of the entrant's product, and is thereby given by

$$(4) \quad P'_M = P'_E = c + \delta.$$

From this price must be deducted the standard distribution margin, so that the manufacturer's price to his distributors is

$$(5) \quad P_M = c + \delta - \gamma.$$

Compare the price to consumers in the presence of exclusive dealing with that charged in its absence. The relevant comparison is with prices set under a high-price strategy, as reported in expression (2). Clearly, Class B customers are worse off in the presence of exclusive dealing, since they would be content to purchase the entrant's product. Class A customers, on the other hand, purchase the original manufacturer's product in either case and could be better off or worse off. They face higher prices with exclusive dealing so long as

$$(6) \quad \delta > \alpha + \gamma,$$

which can be rewritten,

$$(7) \quad \alpha < (\delta - \gamma),$$

and lower prices where these inequalities are reversed.

An economic interpretation of this result is that prices are higher to Class A consumers under exclusive dealing so long as the increased costs of distributing the product alone exceed the brand preference shown for the dominant firm's product. Where this inequality is reversed, however, Class A customers face lower prices in the presence of exclusive dealing.

With a high-price strategy, the manufacturer sets the highest price possible for Class A customers, and accepts the loss of Class B customers. The final price now is the entrant's price plus  $\alpha$ . Since the entrant's resale price is  $(c + \delta)$ , the original manufacturer's price

to consumers is readily determined, and also his price to established distributors:

$$(8) \quad \begin{aligned} P'_M &= c + \alpha + \delta, \\ P_M &= c + \alpha + \delta - \gamma. \end{aligned}$$

At these prices, Class A customers purchase the original manufacturer's product despite a price differential of  $\alpha$ . Class B customers, however, purchase the entrant's product through alternate channels of distribution. In this case, we find that prices to both classes of customers are higher in the presence of exclusive dealing.

The implications of this analysis are striking. Under the specified market conditions, the original manufacturer profits by imposing exclusive dealing requirements on his distributors, regardless of his choice of pricing strategy.

#### IV. Exclusive Dealing with Distributor Choice

In this section, we ask whether the actions of distributors can constrain manufacturer behavior. First, consider the case where brand preferences by Class A consumers are relatively strong, so that

$$(9) \quad \alpha > \delta - \gamma.$$

In this situation a distributor knows that if he did not abide by the exclusive dealing requirements, his sales would be limited to Class B customers. Class A customers would continue to purchase the original manufacturer's product even at a price which encompassed higher distribution costs. Therefore, to go along with this requirement, the distributor must be certain that his profits under exclusive dealing are at least as great as would be earned on sales of the entrant's product to Class B customers alone. The profit comparison is different for the two pricing strategies so we consider them separately.

With a low-price strategy, the original manufacturer retains sales to both classes of customers. Since unit costs and margins are assumed constant, distributor profits depend

on which choice provides the greater volume. The condition for distributor acceptance of exclusive dealing is that

$$(10) \quad Q_B^E(c + \gamma) \leq Q_A^M(c + \delta + \lambda) + Q_B^M(c + \delta + \lambda),$$

where the subscripts refer to sales made to a particular class of customers and the superscripts refer to the manufacturer of the product sold:  $M$  for the original manufacturer and  $E$  for the entrant. The terms in parentheses refer to the prices associated with the indicated quantities.

In this expression, we introduce  $\lambda$  which measures the original manufacturer's markup over costs. The manufacturer selects the highest possible value of  $\lambda$  consistent with distributor choice, which is that value which makes expression (10) an equality. Where the value of  $\lambda$  is positive, the distributor accepts exclusive dealing arrangements imposed by the original manufacturer even at a price higher than the latter's preferred limit-entry price. If so, the constraint imposed by distributor choice is not binding and the limit-entry prices are those stated above. On the other hand, where  $\lambda$  is negative, the distributor choice constraint is binding and the relevant prices are lower by this amount.

The former result seems particularly likely in this case. So long as the number of Class A customers is sufficiently large, the first term on the right-hand side of expression (10) will lead the inequality to hold despite any depressing effect on quantity of higher manufacturer prices.

When the original manufacturer pursues a high-price strategy, a different comparison is relevant. In this case, the distributor is limited to Class A customers if he accepts exclusive dealing and Class B customers otherwise. The condition for distributor acceptance of exclusive dealing is now

$$(11) \quad Q_B^E(c + \gamma) \leq Q_A^M(c + \delta + \alpha + \lambda).$$

Unlike the previous comparison, this inequality is unlikely to be satisfied for positive  $\lambda$ . It depends critically on the relative sizes of

the two classes of customers. Larger numbers of Class B customers will generally lead the distributor to reject exclusive dealing requirements, and as a result, they would not be imposed.

Consider the special case where there are equal numbers of the two classes of customers. The aggregate demand schedules of each class are here identical except that the demand for the original manufacturer's product by Class A customers lies above that of Class B customers by the amount  $\alpha$ . In this case, the inequality in expression (11) does not hold unless  $\lambda$  is sufficiently negative so that the right-hand side quantity is larger. The constraint imposed by distributor choice is now binding, and the original manufacturer's price is lower than previously determined.

While it is therefore possible for the original manufacturer to secure exclusive dealing arrangements by setting a sufficiently low price, it is not likely to be in his interests to do so. The ability of the original manufacturer to pursue a high-price strategy along with exclusive dealing arrangements is limited by distributor choice.

Under weak brand preferences, the inequality given in expression (9) is reversed, and the original manufacturer's position is weaker than before. Again, the distributor's profits depend on his quantity sold. Where the original manufacturer follows a low-price strategy, the distributor's condition for accepting exclusive dealing requirements is given by

$$(12) \quad Q_A^E(c + \gamma) + Q_B^E(c + \gamma) \\ \leq Q_A^M(c + \delta + \lambda) + Q_B^M(c + \delta + \lambda).$$

Recall that the demand of Class A consumers for the original manufacturer's product is identical to their demand for the entrant's product at a price lower by  $\alpha$ . Therefore, we can compare this demand for the two products by adjusting the prices indicated for the original manufacturer's product. Focus here on the first term on each side of expression (12). Letting  $\lambda$  temporarily be zero, the left-hand side price for Class A

consumers is  $(c + \gamma)$  while the adjusted right-hand side price is  $(c + \delta - \alpha)$ . The latter is greater than the former if

$$(13) \quad \delta - \alpha > \gamma,$$

which implies

$$(14) \quad \alpha < \delta - \gamma.$$

This is precisely the case we are considering. The effective price for Class A customers is therefore higher, the quantity smaller, and the inequality is not supported by these consumers. The same result applies for Class B customers, where the higher price on the right-hand side leads to smaller quantities, which again is contrary to the indicated inequality and thereby to the acceptance of exclusive dealing.

To this point, our comparisons assume that prices are unchanged from limit-entry values so that  $\lambda$  is zero. Exclusive dealing, however, will be accepted at some negative value of  $\lambda$ . In this case, distributor choice imposes a binding constraint on manufacturer decisions, and prices must be lowered from limit-entry values to achieve exclusive dealing.

The corresponding comparison for a high-price strategy in the presence of weak brand preferences is given by

$$(15) \quad Q_A^E(c + \gamma) + Q_B^E(c + \gamma) \\ \leq Q_A^M(c + \delta + \alpha + \lambda).$$

So long as there are a substantial number of Class B consumers, this inequality is unlikely to be satisfied, and exclusive dealing will not be practiced. Again, distributor choice imposes a binding constraint on the decisions of the original manufacturer.

We conclude that exclusive dealing is more likely where brand preferences are strong. Furthermore, distributors are much less likely to abide by exclusive dealing where high-price strategies are followed. Low-price strategies which attract both classes of consumers to established distributors and to the original manufacturer are a far more likely outcome.



### V. Exclusive Dealing for Entry Deterrence

In this section, we consider the conditions under which actual prices may exceed the limit-entry prices reported above. What is relevant here is the ability of the original manufacturer to make a credible threat or commitment to respond strategically to a new rival and thereby charge higher prices.

With exclusive dealing requirements, the original manufacturer can exclude rivals with limit-entry prices that exceed costs, which is not otherwise possible. Moreover, if the original manufacturer can commit himself to this price if entry occurs and can communicate this commitment to prospective entrants, he can charge more than the limit-entry price.<sup>4</sup> While this type of action is of course independent of exclusive dealing, that policy increases its feasibility.

With exclusive dealing, the original manufacturer's limit-entry price provides positive profits while it offers no profits otherwise. The threat to contest vigorously the entry of a new manufacturer is now more plausible. Exclusive dealing thereby increases the credibility of the threat to exclude rivals and contributes to entry deterrence.

The price that the original manufacturer actually sets depends on the feasibility of lowering his price at the onset of entry. Where a successful commitment can be made, he is *not* limited to the limit-entry prices reported above. For this reason, *it is the presence of exclusive dealing rather than the specific limit-entry prices reported above which is critical for entry deterrence.*

### VI. An Application

In 1977, C. R. Laurence Co. filed suit in Los Angeles against the General Electric Company alleging, among other things, that the latter had attempted to impose exclusive dealing requirements and had cut off all sales

when these requirements were not followed.<sup>5</sup> That case provides a useful example of the circumstances in which exclusive dealing arrangements may be imposed by manufacturers as well as the anticompetitive consequences which may follow.

The suit concerned silicone sealants, a new and greatly advanced product used in the glazing industry. The leading U.S. manufacturer of silicone sealants was the General Electric Company, whose national market share reached about 75 percent during the mid-1970's. The product had been developed and patented by the French chemical company, Rhone-Poulenc, and G.E. was an early licensee.

Silicone sealants are sold through distributors who sell many other products as well. C. R. Laurence was a large distributor, and accounted for between 7 and 9 percent of total G.E. sales of sealant during this period. Although silicone sealants represented only about 2 to 3 percent of total C. R. Laurence sales, they were highly visible since most customers require some form of sealant. Many customers therefore pay particular attention to its price.<sup>6</sup>

As an early product on the U.S. market, the G.E. brand acquired a strong following, and many purchasers were willing to pay a substantial premium. Others, however, sought the brand with the lowest price.

There appear to be substantial economies in distributing sealant in the same process and location as other glazing supplies. Laurence provides an inventory of 15,000 different products which buyers purchase and use in conjunction with sealant. Distribution of sealants through other outlets would therefore be inefficient.

<sup>4</sup>For discussion of the importance of threats and commitments for oligopoly behavior, see Lester Telser, (1966) and Salop (1979b).

<sup>5</sup>*C. R. Laurence Co. v. General Electric Co.* (1977). All of the information presented in this section was obtained from the complaint in this matter or from interviews with Donald E. Friese, Vice-President and General Manager, C. R. Laurence Co. on October 6 and 20, 1982. Both the complaint and a record of these interviews is available from the authors on request. No data obtained from General Electric or its competitors during the course of discovery in this matter were used in the preparation of this paper.

<sup>6</sup>Donald Friese (1982).

In addition to economies of scope, there appear to be scale economies as well. C. R. Laurence had gained a leading position throughout California with a market share of between 30 and 35 percent. No other distributor was even one-third as large. With this scale of operation, Laurence provided various delivery and credit services to "good" customers. Moreover, its distribution margin was sufficient to generate a substantial return and make it a highly profitable operation.

In the mid-1970's, Rhodia sought to enter the U.S. market for silicone sealants with a highly comparable product that C. R. Laurence believes to be "almost exactly the same" (see Friese). Rhodia is the American subsidiary of the same French chemical company, Rhone-Poulenc, which had developed this product and manufactured it in Europe for many years. Moreover, its price to distributors was much lower than G.E.'s price.

General Electric learned that C. R. Laurence was distributing Rhodia products in late 1975. Even though the latter had been a leading distributor of silicone sealants, G.E. terminated Laurence's distributorship on June 1, 1976. Throughout the industry, it was believed that C. R. Laurence was dropped as an authorized distributor of G.E. products precisely because it started to carry Rhodia products. Apparently, G.E. was willing to eliminate even one of its leading distributors if a rival product was handled. Subsequently, at least through the period of the litigation, no other major distributor of G.E. products also sold the Rhodia brand.

The antitrust action which followed emphasized the anticompetitive effects of imposing exclusive dealing arrangements in this industry. Without the availability of distributors like C. R. Laurence in many market areas, Rhodia found it more difficult and costly to sell in the United States. Exclusive dealing arrangements were apparently used to raise the costs of entry at the manufacturing stage of production, and to maintain G.E.'s dominant market position.

## VII. Conclusions

The current state of opinion towards vertical arrangements such as exclusive dealing

has shifted dramatically from one extreme to the other. Rather than being considered as anticompetitive, they are now frequently viewed as procompetitive.

In this regard, we recall Thomas Kuhn's statement that "there are losses as well as gains in scientific revolutions, and scientists tend to be peculiarly blind to the former" (1970, p. 167). The emergence of a new position as to the competitive implications of vertical arrangements represents a scientific revolution in the economics of market relationships, and we need to be careful not to go too far. In regard to exclusive dealing, we are in danger of doing just that.

The policy conclusions which follow from our analysis are evident. Unless one posits prohibitive costs of judicial determination, exclusive dealing arrangements should not be viewed as per se illegal, but neither should they be viewed as per se legal. A full analysis of the market conditions under which they are practiced is required to determine whether there are anticompetitive effects. Moreover, this analysis does not appear unusually complex or difficult by antitrust standards. Our identification of a plausible set of circumstances where exclusive dealing arrangements have anticompetitive implications suggests an antitrust standard of the "rule of reason."

## REFERENCES

- Director, Aaron and Levi, Edward H., "Law and the Future: Trade Regulation," *Northwestern University Law Review*, May-June 1956, 51, 281-96.
- Friese, Donald E., Interviews, October 6 and 20, 1982.
- Hirshleifer, Jack, *Price Theory and Applications*, 2d ed., Englewood Cliffs: Prentice Hall, 1980.
- Kuhn, Thomas S., *The Structure of Scientific Revolutions*, 2d ed., Chicago: University of Chicago Press, 1970.
- Marvel, Howard P., "Exclusive Dealing," *Journal of Law and Economics*, April 1982, 25, 1-25.
- Salop, Steven C., (1979a) "Monopolistic Competition with Outside Goods," *Bell Journal*

- of *Economics*, Spring 1979, 10, 141-56.
- \_\_\_\_\_, (1979b) "Strategic Entry Deterrence," *American Economic Review Proceedings*, May 1979, 69, 335-38.
- Scherer, F. M., *Industrial Market Structure and Economic Performance*, 2d ed., Chicago: Rand McNally, 1980.
- Telser, Lester G., "Cut-throat Competition and the Long Purse," *Journal of Law and Economics*, October 1966, 9, 259-77.
- Willig, Robert, "Multiproduct Technology and Market Structure," *American Economic Review Proceedings*, May 1979, 69, 346-51.
- C. R. Laurence Co. v. General Electric Co.*, U.S. District Court, Central District of California, Complaint, January 7, 1977.

# Enlistments in the All-Volunteer Force: Note

By CHARLES DALE AND CURTIS GILROY\*

In their 1983 article on the determinants and forecasts of military personnel supply, Colin Ash, Bernard Udis, and Robert McNown (A-U-M) provide evidence of "...lower pay elasticities than had been previously estimated, [and] no significant effect of unemployment on recruitment..." (p. 147). These findings are derived from a model which first appeared in McNown et al. (1980). While the forecasting accuracy of their analysis is encouraging, we have also been able to predict future military accessions with a very different set of estimating equations, a variant of those found in our earlier article (1983). These equations are quite robust in yielding both sizable and significant pay and unemployment elasticities.

In Section I, we specify a time-series model from which are derived relatively large pay and unemployment effects, and we compare these results to those obtained by A-U-M. In Section II, this model is used to predict future Army enlistments of nonprior service male high school graduates.

## I. An Alternative Time-Series Model

Ash et al. failed to find an unemployment effect on enlistments because they did not have available the most appropriate data for estimating their equations. They used data on accessions and estimated their model over the draft and postdraft eras (20 semiannual observations from 1967: II through 1976: II). We benefited from special monthly data

tabulations on enlistment contracts signed provided to us by the Defense Manpower Data Center, and estimated our model over the all-volunteer force period (78 monthly observations from 1975 (10) through 1982 (3)).

Although A-U-M properly acknowledged that their theoretical model is formulated on applications to enlist, or "contracts," their empirical analysis is based on accessions (p. 146). We will show here that our use of contracts data produces results that are more in agreement with both economic theory and a priori expectations.

We agree with A-U-M that contracts, rather than accessions, are the correct data to use. Contracts are supply determined—individuals may sign contracts to enlist now, but actually begin their enlistment periods up to a year later. Accessions, on the other hand, are more demand determined by recruiters, since recruiters normally have three-month quotas to fill. Thus, contracts data are the appropriate type to use for estimating military supply equations. Although we have estimated equations for each of the four branches of the armed services as well as for white and nonwhite racial groups, we limit our analysis here to the Army model.

We have not attempted to replicate the A-U-M study exactly, for several reasons. First, they used *all* male Army accessions as the dependent variable, while we estimate only enlistment contracts of *high school graduates*, the group which is of most interest to the Army. Second, they used semiannual observations, which helped to smooth out some of the seasonal fluctuations in the data, but we are better able to account for seasonality with a quarterly dummy variable. Finally, we have included important educational variables in our equations; attempting to add them to A-U-M's equations leads to severe multicollinearity problems. Since our central point is that enlistment contracts, rather than accessions, are the proper data to

\*Research Economist and Chief Economist, respectively, U.S. Army Research Institute, 5001 Eisenhower Avenue, Alexandria, VA 22333. We are grateful to Larry Holmes for helpful discussions, and to Philip Knorr, Cavan Capps, and John Nagel for research assistance. The views expressed are our own and not necessarily those of any of the aforementioned individuals, the U.S. Army Research Institute, or the Department of Defense.

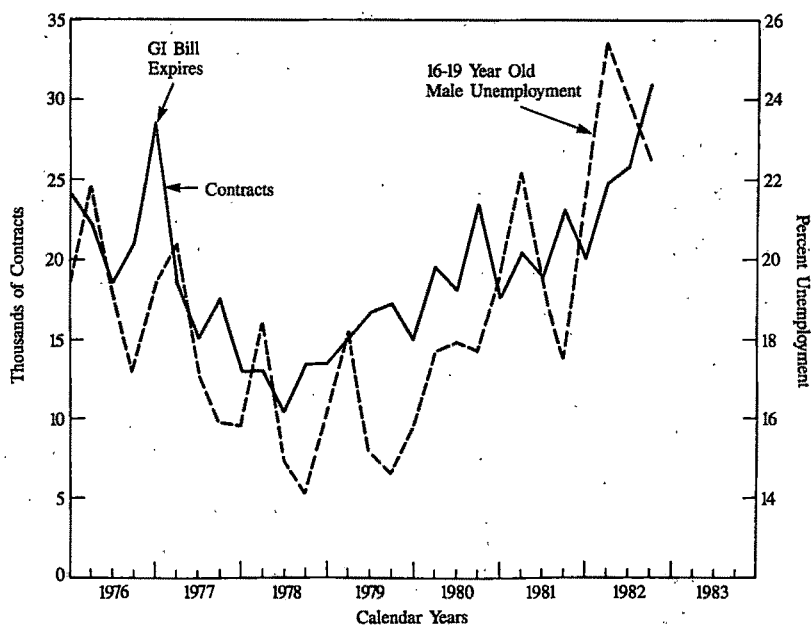


FIGURE 1. CONTRACTS OF ARMY MALE, NONPRIOR SERVICE HIGH SCHOOL GRADUATES ARE CLOSELY CORRELATED WITH UNEMPLOYMENT RATES

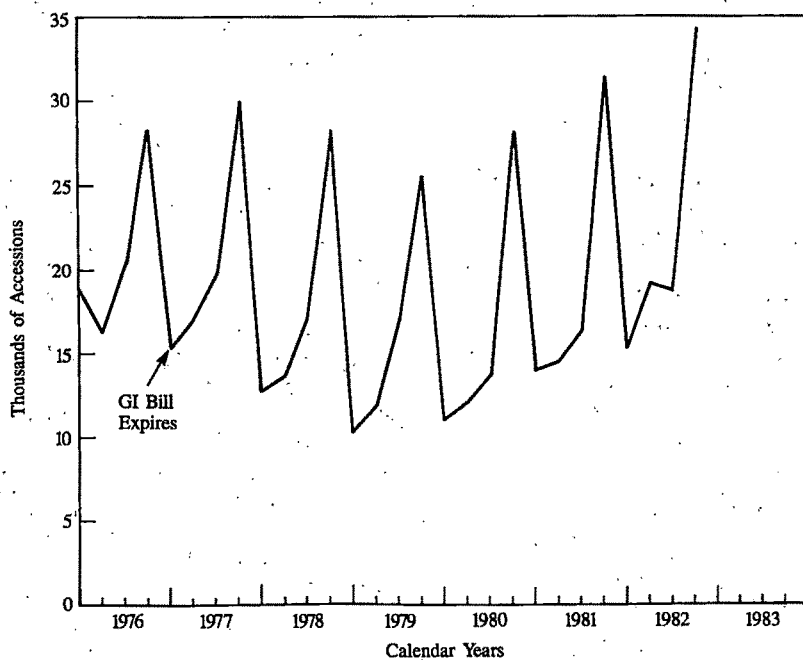


FIGURE 2. ACCESSIONS OF ARMY MALE, NONPRIOR SERVICE HIGH SCHOOL GRADUATES SHOW A STRONG SEASONAL PATTERN

TABLE 1—GLS REGRESSION RESULTS FOR ENLISTMENTS OF ARMY MALE  
NONPRIOR SERVICE HIGH SCHOOL GRADUATES  
MONTHLY DATA, OCTOBER 1975 TO SEPTEMBER 1982

Variable	Lagged Wage Term		Leading Wage Term	
	Accessions (a)	Contracts (b)	Accessions (c)	Contracts (d)
Intercept	3963.8 (3.2) <sup>b</sup>	207.3 (0.3)	-5027.3 (-4.5) <sup>b</sup>	-2637.1 (-4.1) <sup>b</sup>
UM	74.6 (5.4) <sup>b</sup>	34.7 (4.8) <sup>b</sup>	3.1 (0.2)	18.3 (2.7) <sup>b</sup>
W-1	-102.2 (-3.8) <sup>b</sup>	0.2 (0.0)		
W+4			118.8 (5.7) <sup>b</sup>	58.5 (4.8) <sup>b</sup>
GI	-271.2 (-1.2)	755.2 (6.9) <sup>b</sup>	-225.9 (-1.1)	824.2 (7.3) <sup>b</sup>
BILL	-305.0 (-0.9)	-36.0 (-0.2)	484.9 (1.6)	407.0 (2.3) <sup>a</sup>
VEAP	-506.6 (-0.9)	-205.0 (-0.7)	842.7 (1.6)	527.9 (1.7) <sup>a</sup>
KICK	-5.0 (-0.9)	4.0 (1.3)	0.3 (.06)	7.8 (2.7) <sup>b</sup>
TARGET	-224.6 (-2.1)	27.9 (0.5)	82.0 (.84)	158.6 (2.8) <sup>b</sup>
Q3	581.2 (9.2) <sup>b</sup>	170.1 (5.1) <sup>b</sup>	409.0 (5.9) <sup>b</sup>	77.5 (2.0) <sup>a</sup>
R <sup>2</sup>	.67	.75	.74	.81

Source: Data from Defense Manpower Data Center and U.S. Bureau of Labor Statistics.

Notes: *t*-ratios are shown in parentheses. Dependent Variable is  $A/P$  where  $A$  = Army accessions or contracts of male nonprior service high school graduates and  $P$  = male population of 16-19-year-olds. Independent Variables are  $UM$  = unemployment rate of all males 16-19-years-old,  $W-1$  = ratio of first-year military compensation (including basic pay, allowances for quarters and subsistence, and the federal tax advantage) to average weekly earnings in the private economy, with a 1-month lag,  $W+4$  = wage ratio with a 4-month lead,  $GI$  = dummy variable = 1 in December 1976, when the GI bill expired, and = 0 everywhere else,  $BILL$  = Maximum monthly benefit for a GI bill beneficiary without dependents, deflated by the Consumer Price Index. Variable is set = 0 after December 1976,  $VEAP$  = maximum monthly benefit for a VEAP beneficiary without dependents, deflated by the consumer price index. Variable is set = 0 before January 1977,  $KICK$  = maximum value of "kicker" payments—amounts over and above VEAP benefits to those enlisting in designated military specialties—deflated by the Consumer Price Index,  $TARGET$  = binary variable = 1 from November 1979 to August 1981 (a period characterized by special recruiting policies targeted to the high school graduate population), and = 0 everywhere else,  $Q3$  = seasonal dummy = 1 in July, August and September, and = 0 everywhere else.

<sup>a</sup>Significant at .05 level.

<sup>b</sup>Significant at .01 level.

use in studies of this type, the exact form of the equations used is not crucial to our argument.

The distinction between contracts and accessions data is extremely important and may be seen in Figures 1 and 2. In Figure 1, the seasonally unadjusted 16 to 19-year-old

male unemployment rate is superimposed on male Army contract data. The correlation is certainly not perfect, but there is, nonetheless, a visible relation between unemployment and contracts signed. In contrast, the accessions data in Figure 2 clearly show how the expected seasonal pattern predominates,

as most high school graduates access in the summer, regardless of when they sign their contracts.

An equally dramatic difference between our contracts data and the A-U-M accessions data may be seen by considering the December 1976 expiration of the GI bill. The GI bill was a noncontributory educational program which was replaced in January 1977 by the less generous Veterans' Educational Assistance Program (VEAP). There was, consequently, a tremendous rush to sign contracts in December, 1976—a classic economic supply response—as shown in Figure 1. The actual accessions, however, were very low in December 1976, as Figure 2 shows, since relatively few people begin their enlistment tours in December. Again, contracts data rather than accessions data are shown to be the more appropriate series for testing this supply model.

In order to make meaningful comparisons of results from using identical equations with accessions data and contracts data, the 1975–82 time period was chosen—years during which both the all-volunteer force (*AVF*) was in existence and the requisite data were available. Limiting the analysis to the *AVF* period avoided the problem of mixing up the draft and postdraft years which required A-U-M to incorporate a "draft threat" variable in their estimating equation. Using only *AVF* years permits a purer test of the supply response.

The variables included in the equation were the male teenage unemployment rate (*UM*); the ratio of military pay to civilian earnings (*W*); three variables to measure educational benefits (*GI*, *BILL*, and *KICKER*); recruiter effort (*TARGET*); and a seasonal dummy (*Q3*). These variables are described in Table 1.

Roughly approximating A-U-M, we estimated equations with a lagged relative pay rate term (Table 1, cols. a and b), but did not obtain reasonable results with either accessions or contracts data. The equation using accessions data produced many "wrong" signs, and the contracts data equation produced statistically insignificant results for most variables.

TABLE 2—PAY AND UNEMPLOYMENT ELASTICITIES FOR ARMY ENLISTMENT RATES

Study	Elasticities	
	Pay	Unemployment
Accessions Data:		
McNown et al. (p. 121)	.816	.045
A-U-M (p. 153)	.881	.133
Contracts Data:		
D and G (p. 49) <sup>a</sup>	2.3	.94
D and G (Table 1, col. d) <sup>b</sup>	3.9	.81

<sup>a</sup> Our earlier paper.

<sup>b</sup> This paper.

The preferred model is the one that appears in columns c and d in Table 1. The distinguishing feature of these equations is the four-month lead on the pay variable—not attempted until the present work, but which was suggested by Harry Gilman (1970). A four-month lead produced a better statistical fit than three or five month leads.

Columns c and d of Table 1 clearly show the effect of using contracts data vs. accessions data. The economic supply response is shown by the unemployment term which is large and statistically significant in the contracts equation, a result one would expect from both Figure 1 and with what recruits themselves are reporting; it is insignificant in the accessions equation. Indeed, in a recent survey of enlistees into the Army, over 40 percent mentioned unemployment as a factor (Timothy Elig et al., 1982). The supply response expected by the expiration of the GI bill in December 1976 is strongly picked up by the variable *GI* in the preferred contracts equation, but that variable even has the wrong sign in the accessions equation. All other education variables in the contracts equation are also statistically significant and display the correct signs. Finally, the variable *W* + 4 shows that relative pay is also an important factor—in both size and significance—in the enlistment decision.

Table 2 summarizes the differences in pay and unemployment elasticities from equations estimated with accessions and contracts data. Both the unemployment and pay terms show up much more clearly in the contracts equations.

TABLE 3—FORECASTS OF ARMY ENLISTMENTS OF MALE NONPRIOR SERVICE HIGH SCHOOL GRADUATES  
(Thousands of Contracts)

	FY81	FY82	FY83	FY84	FY85	FY86	FY87
D and G (1983)	78.1	97.1	95.9	82.6	72.7	66.5	61.9
D and G (Present Work)	75.8	94.0	87.5	75.7	66.8	59.5	52.4
Actual	79.9	95.6					

Note: FY82 actual is through June 1982 at an annual rate; FY87 projections are 3-month projections at annual rates.

## II. Forecasts of Army Enlistments

Ash et al. correctly point out that forecasting accuracy is an important criterion for the usefulness of an econometric model, so we report our projections of Army enlistments in Table 3. We assumed military pay will be frozen in fiscal year 1984, and military and civilian pay will increase at 6 percent per year from fiscal year 1985 onward. Our earlier work shows more optimistic projections because we did not anticipate any pay freezes. The overall trend in Army enlistments is clearly downward, largely due to a decreasing youth population and projections of declining unemployment rates.

Predicted Army enlistments from both our earlier paper and the present work are close to the actual values, differing by between 2–5 percent in fiscal year 1981 and about 2 percent in fiscal year 1982. It is not possible, however, to compare our forecasts with those of A-U-M since they did not provide projections beyond 1979.

We conclude by noting that we agree with A-U-M's conclusion that "current military pay policy might be inadequate to meet manpower requirements in the 1980's" (p. 154). Our results strongly support the view that it is important to maintain the comparability of military pay relative to civilian pay,

especially in the face of both unfavorable demographic trends and an improving economy.

## REFERENCES

- Ash, Colin, Udis, Bernard and McNown, Robert F., "Enlistments in the All-Volunteer Force: A Military Personnel Supply Model and its Forecasts," *American Economic Review*, March 1983, 73, 145–55.
- Dale, Charles and Gilroy, Curtis, "The Effects of the Business Cycle on the Size and Composition of the U.S. Army," *Atlantic Economic Journal*, March 1983, 11, 42–53.
- Elig, Timothy W., Gade, Paul A., and Shields, Joyce L., "1982 Department of the Army Survey of Personnel Entering the Army," manuscript, presented at the meeting of the Military Testing Association, San Antonio, Texas, November 1982.
- Gilman, Harry J., "The Supply of Volunteers to the Military Services," in *The Report of the President's Commission on an All-Volunteer Armed Force*, Vol. 1, Washington: USGPO, 1970.
- McNown, Robert F., Udis, Bernard and Ash, Colin, "Economic Analysis of the All-Volunteer Force," *Armed Forces and Society*, Fall 1980, 7, 113–32.



# Is There an Operational Interest Rate Rule?

By MICHAEL DOTSEY\*

In his 1983 paper, Jeremy Siegel derives a seemingly implementable policy rule involving optimal responses to interest rates. The existence of such a rule would be of tremendous interest to central banks whose monetary policies place heavy weight on responses to interest rates. The Siegel rule is especially appealing because it is (i) an optimal combination policy in the sense of William Poole (1970), and (ii) the proposed implementation of the rule does not require detailed knowledge of the structure of the economy. All that is required is a calculation of the covariance between innovations in prices and interest rates.

Within the confines of a rational expectations equilibrium model, in which Siegel assumes agents do not make use of information embodied in the current nominal interest rate, he is able to design an optimal combination policy that does not require detailed information about the economy. That such an optimal policy exists is not new, but that it can be easily implemented is novel.<sup>1</sup> The policy rule depends solely on the covariance between innovations in the aggregate price level and innovations in the nominal rate of interest, normalized by the variance of innovations in the interest rate. When this index is zero, policy has been set optimally. When the index is positive the feedback term on interest rates in the money supply rule is too large, and when the index is negative the feedback term is too small. Given that one can obtain reduced-form expressions for prices and interest rates, the index is easily computed.

Unfortunately, Siegel's proposal violates Robert Lucas's (1976) critique. That is, he implicitly treats as invariant certain aspects of economic behavior that will generally change when one moves to an "operational interest rate rule." This note shows in detail that in a model where prices are flexible and agents observe local market prices (i.e., the model at least employed verbally by Siegel), that the coefficients in the aggregate supply and demand functions are not invariant to the form of the money supply rule. This lack of invariance will cause Siegel's rule to be nonoperational. The sensitivity of parameters in aggregate supply functions to policy is not restricted to equilibrium models with flexible prices. This property also extends to contracting models with endogenous indexing (see, for example, Jo Anna Gray, 1976). Therefore, Siegel's rule will not be implementable in a wide variety of commonly used macro models.

## I. Basic Micro Structure

In Siegel's model, agents form expectations conditioned on their observation of local prices. In this section a foundation consistent with the aggregate economy presented in Siegel is constructed. Let supply and demand in market  $z$  be represented by

$$(1) \quad y_t^s(z) = a_s(p_t(z) - E p_t | I_t(z)), \quad a_s > 0,$$

$$(2) \quad y_t^d(z) = a_d(r_t + p_t(z) - E p_{t+1} | I_t(z)) \\ + \varepsilon_t^d + z_t, \quad a_d < 0,$$

where  $y$  and  $p$  are logarithms of output and prices, and  $r$  is the nominal rate of interest. The letter  $z$  indicates the  $z$ th market and the notation  $E p_t | I_t(z)$  indicates the expectation of the log of the aggregate price level conditioned on current local information. This information is assumed to include the local price  $p_t(z)$ , past local and aggregate prices,

\*Federal Reserve Bank of Richmond, Richmond, VA 23261. I thank Marvin Goodfriend, Robert King, and Tony Kuprianov for helpful comments. The views expressed in this paper are not necessarily those of the Federal Reserve Bank of Richmond or the Federal Reserve System and are solely my responsibility.

<sup>1</sup>For example see Matthew Canzoneri et al. (1983), and Geoffrey Woglom (1979).

past interest rates, past levels of the money stock, and the values of all past disturbances. As in Siegel, the assumption that it is too costly for agents to process the information contained in the nominal interest rate is maintained.<sup>2</sup> The aggregate disturbance  $\varepsilon_t^A$  is a trendless random walk equal to  $\varepsilon_{t-1}^A + \tilde{\varepsilon}_t^A$ , while  $z_t$  is a local disturbance and is assumed to be white noise.

## II. The Macro Structure

Averaging across markets yields aggregate supply and demand curves similar to those presented by Siegel. (A lagged output term in (1') is omitted for simplicity.)

$$(1') \quad y_t^s = a_s(p_t - \overline{E}p_t|I_t(z)),$$

$$(2') \quad y_t^d = a_d(r_t + p_t - \overline{E}p_{t+1}|I_t(z)) + \varepsilon_t^A.$$

Here the overbar indicates the average of all local expectations and it is assumed that the relative disturbances cancel out across markets.<sup>3</sup>

The model is completed by appending Siegel's equation (3), which describes the money market:<sup>4</sup>

$$(3) \quad m_t = m_0 + m_r r_t = \bar{p}_t + l_y y_t + l_r r_t + \varepsilon_t^L, \\ l_y > 0, l_r < 0.$$

The log of money supplied,  $m_t$ , is determined by a combination policy and is equal to money demand. The disturbance term  $\varepsilon_t^L$  is also a trendless random walk depicted by  $\varepsilon_t^L = \varepsilon_{t-1}^L + \tilde{\varepsilon}_t^L$ .

<sup>2</sup>It is well known that if agents use the nominal interest rate, then interest rate feedback rules have no real effects. For a more detailed discussion of this point see my article with Robert King (1983), or Siegel. The assumption that agents do not use the information contained in the interest rate is employed in an effort to be consistent with the main elements of Siegel's analysis.

<sup>3</sup>The above aggregate supply and demand curves are similar to those derived by myself and King.

<sup>4</sup>For expositional ease it is assumed that the pure money stock component of the money supply rule is constant.

The analysis in Siegel proceeds by examining the innovations in prices and interest rates. For notational convenience, let a tilde over a variable indicate an innovation in that variable (i.e.,  $\tilde{x}_t = x_t - Ex_t|I_{t-1}^*$ ), and where  $I_{t-1}^*$  equals  $I_{t-1}(z)$  plus knowledge of  $\varepsilon_{t-1}^A$  and  $\varepsilon_{t-1}^L$ . However, before proceeding, terms such as  $\overline{E}p_t|I_t(z)$ , and  $\overline{E}p_{t+1}|I_t(z)$  must be analyzed.

## III. Expectations and Information

Equations (1'), (2'), and (3) present a model of an economy that is log linear. Therefore, one may postulate a linear solution for  $p_t(z)$ ,  $p_t(z) = \pi_1 m_0 + \pi_2 \varepsilon_{t-1}^A + \pi_3 \varepsilon_{t-1}^L + \pi_4 z_t + \pi_5 \tilde{\varepsilon}_t^A + \pi_6 \tilde{\varepsilon}_t^L$ . The expectations of the aggregate price level and next period's price level can be determined by the following equations:<sup>5</sup>

$$(4a) \quad Ep_t|I_t(z) = Ep_t|I_{t-1}^* \\ + \beta(p_t(z) - Ep_t|I_{t-1}^*),$$

$$(4b) \quad \overline{E}p_t|I_t(z) = Ep_t|I_{t-1}^* \\ + \beta(p_t - Ep_t|I_{t-1}^*),$$

$$(5a) \quad Ep_{t+1}|I_t(z) = Ep_{t+1}|I_{t-1}^* \\ + \gamma(p_t(z) - Ep_t|I_{t-1}^*),$$

$$(5b) \quad \overline{E}p_{t+1}|I_t(z) = Ep_{t+1}|I_{t-1}^* \\ + \gamma(p_t - Ep_t|I_{t-1}^*).$$

In the above expressions,  $Ep_t|I_{t-1}^* = \pi_1 m_0 + \pi_2 \varepsilon_{t-1}^A + \pi_3 \varepsilon_{t-1}^L = Ep_{t+1}|I_{t-1}^*$ . Using equations (1') and (2'), along with (4b) and (5b), the aggregate supply and demand curves can be expressed as

$$(1'') \quad y_t^s = a_s(1 - \beta)(p_t - Ep_t|I_{t-1}^*),$$

$$(2'') \quad y_t^d = a_d[r_t + p_t - Ep_{t+1}|I_{t-1}^* \\ - \gamma(p_t - Ep_t|I_{t-1}^*)] + \varepsilon_t^A.$$

<sup>5</sup>For more detail, see my article with King, or Thomas Sargent (1979).

Notice that the elasticities of supply and demand are not independent of policy parameters, since the regression coefficients  $\beta$  and  $\gamma$  are composed of all the underlying parameters of the model as well as the variances and covariances of the disturbance terms. This observation represents the essential difference between the model presented in Siegel and the one derived here. Of crucial importance is that this model is derived from the underlying equilibrium conditions in local markets and is therefore consistent with the description of the economy advanced by Siegel.

#### IV. Siegel's Index

The derivation of Siegel's index proceeds from an examination of the innovations in prices and interest rates. Using (1''), (2''), and (3) these may be expressed as

$$(6) \quad \tilde{y}_t = a_s(1-\beta)\tilde{p}_t,$$

$$(7) \quad \tilde{r}_t = - \left[ [1 + l_y a_s(1-\beta)] \tilde{\epsilon}_t^A + [a_s(1-\beta) - a_d(1-\gamma)] \tilde{\epsilon}_t^L \right] / \left[ [a_s(1-\beta) - a_d(1-\gamma)] [l_r - m_r] + a_d[1 + l_y a_s(1-\beta)] \right],$$

$$(8) \quad \tilde{p}_t = \left[ (l_r - m_r) \tilde{\epsilon}_t^A - a_d \tilde{\epsilon}_t^L \right] / \left[ [a_s(1-\beta) - a_d(1-\gamma)] [l_r - m_r] + a_d[1 + l_y a_s(1-\beta)] \right].$$

These latter three expressions are analogous to Siegel's equations (4)–(6), and would be identical if  $\beta = 0$  and  $\gamma = 1$ . This last condition is impossible given the derivations for  $\beta$  and  $\gamma$ .<sup>6</sup>

The goal of monetary policy is to minimize the variance of output around its full-

information value, which in Siegel's context is equivalent to minimizing  $\sigma_p^2$ .<sup>7</sup> This is done by choosing the appropriate value of  $m_r$ , call it  $m_r^*$ . One can also find the value of  $m_r$ , call it  $\hat{m}_r$ , that forces the  $\text{cov}(\tilde{p}, \tilde{r}) = 0$ . In Siegel's model, it turns out that these two values of  $m_r$  are the same (i.e.,  $m_r^* = \hat{m}_r$ ), and that a monetary policy which sets  $\text{cov}(\tilde{p}, \tilde{r}) = 0$  is optimal. Here it can be shown that unless  $\partial\beta/\partial m_r = \partial\gamma/\partial m_r = 0$  when evaluated at  $\hat{m}_r$ , that this is not the case. Therefore, the optimal value of this covariance will end up being a complicated function of the parameters of the model, and Siegel's procedure does not represent an improvement on existing literature.

To see that this is true, it is necessary to carefully examine the expressions for  $\beta$  and  $\gamma$ , as well as the  $\text{cov}(\tilde{p}, \tilde{r})$ . In the following analysis it is assumed that all the disturbances are independent. Using the undetermined coefficients solution for  $p_t(z)$ ,  $\beta$  and  $\gamma$  can be expressed as

$$\beta = \frac{\pi_5^2 \sigma_A^2 + \pi_6^2 \sigma_L^2}{\pi_4^2 \sigma_z^2 + \pi_5^2 \sigma_A^2 + \pi_6^2 \sigma_L^2},$$

$$\gamma = \frac{\pi_2 \pi_5 \sigma_A^2 + \pi_3 \pi_6 \sigma_L^2}{\pi_4^2 \sigma_z^2 + \pi_5^2 \sigma_A^2 + \pi_6^2 \sigma_L^2},$$

where  $\sigma_z^2$ ,  $\sigma_A^2$ , and  $\sigma_L^2$  are the variances of  $z$ ,  $\tilde{\epsilon}^A$ , and  $\tilde{\epsilon}^L$ , respectively. Using the solutions for the  $\pi$ 's, these expressions can be manipulated to yield

$$(9) \quad \beta = \left[ [a_s(1-\beta) - a_d(1-\gamma)]^2 \times [(l_r - m_r)^2 \sigma_A^2 + a_d^2 \sigma_L^2] \right] / \left[ \delta^2 \sigma_z^2 + [a_s(1-\beta) - a_d(1-\gamma)]^2 \times [(l_r + m_r)^2 \sigma_A^2 + a_d^2 \sigma_L^2] \right],$$

$$(10) \quad \gamma = \delta\beta/a_d,$$

<sup>6</sup>If Siegel had postulated an aggregate demand function of the form  $y_t^d = a_d(r_t + p_t - \bar{E}p_{t+1}) + \epsilon_t^A$ , then our results would be identical with  $\beta = \gamma = 0$ .

<sup>7</sup>Since  $\beta$  is now a function of  $m_r$ , this is no longer the case. Also, Robert Barro (1976) makes an appealing case for minimizing the variance of local output around its full-information value. This policy is also not equivalent to the one used by Siegel.

where  $\delta = [a_s(1-\beta) - a_d(1-\gamma)][l_r - m_r] + a_d[1 + l_y a_s(1-\beta)]$ . It is clear that  $0 \leq \beta \leq 1$  and  $\beta \rightarrow 0$  as  $\sigma_z^2 \rightarrow \infty$ , and that  $\beta \rightarrow 1$  as  $\sigma_z^2 \rightarrow 0$ . The  $\text{cov}(\tilde{p}, \tilde{r})$  is given by

$$(11) \quad \text{cov}(\tilde{p}, \tilde{r}) = [a_d[a_s(1-\beta) - a_d(1-\gamma)]\sigma_L^2 - [l_r - m_r][1 + l_y a_s(1-\beta)]\sigma_A^2] / \delta^2.$$

The expressions for  $\beta$ ,  $\gamma$ , and  $\delta$  will now be analyzed to show that setting the  $\text{cov}(\tilde{p}, \tilde{r}) = 0$  is not in general optimal. First observe that the value of  $m_r$  which sets  $\text{cov}(p, r) = 0$  is given by

$$\hat{m}_r = l_r - [a_d[a_s(1-\beta) - a_d(1-\gamma)]\sigma_L^2] / [1 + l_y a_s(1-\beta)]\sigma_A^2.$$

Also, using (8) it can be shown that the optimal value of  $m_r$ ,  $m_r^*$ , satisfies the following equation:

$$(12) \quad (l_r - m_r)\delta^2\sigma_A^2 + \delta[(l_r - m_r)^2\sigma_A^2 + a_d^2\sigma_L^2](\partial\delta/\partial m_r) = 0.$$

The value of  $m_r$  that satisfies (12) will equal  $\hat{m}_r$  if and only if  $\partial\beta/\partial m_r = \partial\gamma/\partial m_r = 0$  at  $\hat{m}_r$ . In the Appendix, this is shown not to be generally true. It will, however, be true in the case where  $\sigma_z^2 = \infty$ , because this implies that agents will not use any local information in forming expectations. Therefore, the aggregate supply and demand elasticities will not be influenced by changes in policy, since it was only through the expectation-generating process that policy parameters influenced these elasticities.

Since  $m_r^* \neq \hat{m}_r$ , the  $\text{cov}(\tilde{p}, \tilde{r})$  evaluated at  $m_r^*$  will involve all the parameters and variances of the model including the coefficients  $\beta$  and  $\gamma$ . Therefore, the monetary authority could not use the zero covariance criterion to choose  $m_r^*$ . Further, the index derived by Siegel that allows for unambiguously correct movements in policy when the  $\text{cov}(\tilde{p}, \tilde{r})$  strays from zero ought to be monotonic in  $m_r$ . Given that this index will involve the expression for  $\beta$  and  $\gamma$ , it is not clear that

this condition will be met. In short, there is no easy way of implementing a combination policy within the framework envisioned by Siegel.

## V. The Problem Extended to Contracting Models

The problems inherent in implementing Siegel's interest rate rule have been discussed in detail within the context of a model where local information is important. However, similar types of problems will occur in contracting models where wages are preset and the contracts involve endogenous indexing (see Gray).<sup>8</sup> The primary reasons for this are (i) that aggregate supply responses depend upon the optimal degree of wage indexation, and (ii) that the degree of wage indexation is sensitive to policy. Therefore the responsiveness of aggregate supply to movements in prices from their expected value, the expectation being conditioned on last period's information, is sensitive to policy.<sup>9</sup> Equivalently, this means that the coefficient in the aggregate supply schedule is not invariant to policy choices. It is this lack of invariance that will produce problems similar to those outlined above.

## VI. Conclusions

This note has shown that one can generally expect the coefficients in an aggregate supply function to be sensitive to monetary policy. This lack of policy invariance is in agreement with the Lucas critique. Importantly, it provides the mechanism that renders Siegel's interest rate rule nonoperational. For Siegel to make his proposal attractive, he must first derive supply and demand schedules that are invariant to his policy rule. Since a large class of currently used models do not display this behavior, any results along this line are of limited interest.

<sup>8</sup> Gray's model does not involve interest rates. I have worked out the results involving feedback terms on prices. The basic argument should not be sensitive to the particular variable to which monetary policy responds.

<sup>9</sup> Note in Gray's model, the certainty equivalent level of prices,  $p_t^*$ , and  $E_{t-1} p_t$  are equivalent.

## APPENDIX

This Appendix evaluates the derivatives of  $\partial\beta/\partial m_r$  and  $\partial\gamma/\partial m_r$ , by using the implicit function theorem. Using the derivations of  $\beta$ ,  $\gamma$ , and  $\delta$ , define the following system of implicit functions:

$$(A1) \quad 0 = F(\beta, \gamma, \delta; m_r, \dots)$$

$$\begin{aligned} &= \beta - [a_s(1-\beta) - a_d(1-\gamma)]^2 \\ &\quad \times [(l_r - m_r)^2 \sigma_A^2 + a_d^2 \sigma_L^2] \\ & / [\delta^2 \sigma_z^2 + [a_s(1-\beta) - a_d(1-\gamma)]^2 \\ &\quad \times [(l_r - m_r)^2 \sigma_A^2 + a_d^2 \sigma_L^2]] \end{aligned}$$

$$(A2) \quad 0 = G(\beta, \gamma, \delta; m_r, \dots) \equiv \gamma - \delta\beta/a_d,$$

$$\begin{aligned} (A3) \quad 0 &= H(\beta, \gamma, \delta; m_r, \dots) \\ &= \delta - [a_s(1-\beta) - a_d(1-\gamma)][l_r - m_r] \\ &\quad - a_d[1 + l_y a_s(1-\beta)]. \end{aligned}$$

Differentiating this system yields

$$\begin{bmatrix} F_\beta & F_\gamma & F_\delta \\ G_\beta & G_\gamma & G_\delta \\ H_\beta & H_\gamma & H_\delta \end{bmatrix} \begin{bmatrix} \partial\beta/\partial m_r \\ \partial\gamma/\partial m_r \\ \partial\delta/\partial m_r \end{bmatrix} = \begin{bmatrix} -F_{m_r} \\ -G_{m_r} \\ -H_{m_r} \end{bmatrix}.$$

It can be shown that  $\partial\beta/\partial m_r = 0$  when

$$\begin{aligned} m_r &= l_r - [\{a_d[a_s(1-\beta) - a_d(1-\gamma)] \\ &\quad + a_d^2\beta[1 + l_y a_s(1-\beta)]\} \sigma_L^2] \\ &\quad / \{[1 + l_y a_s(1-\beta)] \sigma_A^2 \\ &\quad + a_d\beta[a_s(1-\beta) - a_d(1-\gamma)] \sigma_L^2\}. \end{aligned}$$

This will be equivalent to  $\hat{m}_r$  when  $\beta = 0$ , which occurs when  $\sigma_z^2 = \infty$ . One can also use the derivation of  $\gamma$  to observe that  $\partial\gamma/\partial m_r$

$= (\delta\partial\beta)/(\sigma_A^2\partial m_r) + (\beta\partial\delta)/(\sigma_L^2\partial m_r)$ . Evaluating this at  $m_r = \hat{m}_r$  and  $\sigma_z^2 = \infty$  implies that  $\partial\gamma/\partial m_r = 0$ . Therefore, when  $\sigma_z^2 = \infty$ ,  $\partial\beta/\partial m_r = \partial\gamma/\partial m_r = 0$  at  $\hat{m}_r$ , and  $\hat{m}_r = m_r^*$ . However, in general  $\partial\beta/\partial m_r$  and  $\partial\gamma/\partial m_r$  do not equal zero when evaluated at  $m_r$ , implying that  $\hat{m}_r \neq m_r^*$ .

## REFERENCES

- Barro, Robert J., "Rational Expectations and the Role of Monetary Policy," *Journal of Monetary Economics*, January 1976, 2, 1-32.
- Canzoneri, Matthew, Henderson, Dale and Rogoff, Kenneth, "The Information Content of Interest Rates and the Effectiveness of Monetary Policy Rules," *Quarterly Journal of Economics*, November 1983, 98, 545-66.
- Dotsey, Michael and King, Robert, G., "Monetary Instruments and Policy Rules in a Rational Expectations Environment," *Journal of Monetary Economics*, September 1983, 12, 357-82.
- Gray, Jo Anna, "Wage Indexation: A Macroeconomic Approach," *Journal of Monetary Economics*, April 1976, 2, 221-35.
- Lucas, Robert E., Jr., "Econometric Policy Evaluation: A Critique," in K. Brunner and A. H. Meltzer, eds., *The Phillips Curve and Labor Markets*, Carnegie-Rochester Conference Series on Public Policy, Vol. 1, *Journal of Monetary Economics*, Suppl., 1976, 62, 19-46.
- Poole, William, "Optimal Choice of Monetary Policy Instruments in a Simple Stochastic Macro Model," *Quarterly Journal of Economics*, May 1970, 84, 197-216.
- Sargent, Thomas, J., *Macroeconomic Theory*, New York: Academic Press, 1979.
- Siegel, Jeremy, "Operational Interest Rate Rules," *American Economic Review*, December 1983, 73, 1102-09.
- Woglom, Geoffrey, "Rational Expectations and Monetary Policy in a Simple Macroeconomic Model," *Quarterly Journal of Economics*, February 1979, 93, 91-105.

## U.S. Monetary Policy and the Exchange Rate: Comment

By STEVEN AMBLER AND RONALD MCKINNON\*

In the March 1984 issue of this *Review*, Henry Goldstein and Stephen Haynes were sharply critical of Ronald McKinnon's (1982) analysis of how the foreign exchanges impinge on the American monetary system. McKinnon contends that central banks in open economies should take pressure in the foreign exchanges into account in formulating domestic monetary policy. Money growth should be higher than normal when the domestic currency tends to appreciate against "hard" foreign monies, and vice versa. By following such a rule, the authorities in any industrial economy (including the United States) can better balance the supply to the (international) demand for the national money so as to prevent unexpected inflations or deflations.

Somewhat strangely, Goldstein-Haynes first try to reject McKinnon's idea that shifts in international portfolio preferences indirectly destabilize the demand for domestic money. Then they go on to propose an "alternative" view that governments often intervene to stabilize their foreign exchange rates, allowing domestic money growth to vary in support, as a means of stabilizing the domestic price level. They cite approvingly the unusually high money growth in Germany in 1977-78, when the foreign exchange value of the deutsche mark was increasing sharply. But this is not an alternative view at all! It is McKinnon's view of how the central bank should respond to a signal from the foreign exchanges in order to prevent sudden exchange overvaluation and deflation.

Indeed, McKinnon criticized the failure of the U.S. Federal Reserve System to respond properly to large fluctuations in the dollar exchange rate from 1970 to mid-1982 when other industrial countries were intervening

—not very successfully—to smooth their exchange rates. The resulting cyclical fluctuations in the rest of the world's money *and* in the dollar exchange rate then fed back into sharp cycles of inflation and deflation within the United States itself.

Let us now turn to Goldstein and Haynes' empirical work measuring international influences on the American economy. The potential importance of the exchange rate was established in the econometric interchange among McKinnon, Christopher Radcliffe, Kong-Yam Tan, Arthur Warga, and Thomas Willett in the December 1984 issue of this *Review*. Fluctuations in the dollar exchange rate appear to be an important leading indicator of subsequent changes in American nominal *GNP* and in the dollar prices of tradable goods—for which estimates are more fully worked out in McKinnon (1984) and Tan (1984).

In contrast, Goldstein and Haynes omitted the dollar exchange rate from their empirical analysis and thus failed to come to grips with this important part of the McKinnon hypothesis. We would welcome further empirical work where they included the (lagged) dollar exchange rate as an explanatory variable in their St. Louis regressions. If floating was "clean," the exchange rate would be a sufficient statistic to measure (indirect) fluctuations in the demand for money in the United States.

Under the world dollar standard and "dirty" floating, however, foreign central banks continually intervene to (partially) stabilize their dollar exchange rates. Consequently, portfolio shifts for or against dollar assets—mainly bonds—can lead to changes in foreign money growth rates which also reflect changes in money market conditions in the United States (McKinnon, 1982). In addition, foreign money growth could well have an independent effect on the dollar prices of American tradable goods. To test this aspect of the McKinnon hypothesis,

\*Department of Economics, Stanford University, Stanford, CA 94305.

Goldstein and Haynes regress U.S. price changes on U.S. and foreign money growth rates. They find that foreign money is not statistically significant. Even here, however, their statistical procedures are inappropriate.

First, McKinnon states that "wholesale indices come closer than consumer prices to provide a common denominator of tradable goods" (1982, p. 332). Somewhat inexplicably, Goldstein and Haynes use instead the U.S. GNP deflator as their dependent variable without warning the reader that they had changed price indices from those McKinnon had originally used. The GNP deflator contains large nontradable components—mainly services—which are less sensitive to foreign influences. In contrast, foreign money growth does have a significant impact on American wholesale (producer) prices, as indicated in the preceding interchange.

Precisely which price index the central bank should aim to stabilize is an open question. A case can be made for focusing on a broad commodities index from which changes in the cost of services, most of which are impossible to measure accurately, are omitted. This case is made even stronger if exchange rate stabilization among hard-currency countries, with similar price objectives, is deemed desirable.

Second, by using twelve-quarter moving averages for their explanatory monetary variables, Goldstein and Haynes tend to suppress—or smooth out—the sharp cyclical fluctuations of one to two years in "world" money growth with which McKinnon was so concerned. Their statistical regression procedures were not set up correctly to test the McKinnon hypothesis.

From our experience so far with floating exchange rates in the 1970's and 1980's, how can we best summarize the impact of the foreign exchanges on the American monetary system? Because of lags in the collection of data on both foreign and American money supplies, the dollar exchange rate is the more immediately relevant leading indicator of inflation or deflation to come in the American economy. In the intermediate run, however, growth in foreign hard monies should be monitored by the Federal Reserve in order to prevent undue expansion or contraction in

the joint monetary base of the system as a whole.

By this international standard, how well has the Fed done in 1984? Unfortunately the Fed erred by keeping U.S. M1 growth at a normal annual rate of 5 to 6 percent *despite* the further sharp appreciation of the dollar exchange rate from a level which was already grossly overvalued. This international portfolio shift in favor of a broad spectrum of dollar assets signalled that 1) the derived demand for U.S. base money had increased, and 2) that the brunt of the resulting deflationary pressure in 1985 would be borne by American export and import-competing industries—resulting in a further upsurge of protectionist sentiment in the United States.

What about growth in "world" money in the first nine months of 1984? To prevent their currencies from depreciating even further against the dollar and provoke more protectionism in the United States, both the German Bundesbank and the Bank of Japan were forced to curtail growth in their own M1s below normal. Consequently joint money growth in the three countries was too low, and reinforced the signal given by the appreciating dollar of too much deflationary pressure—with the risk of an unnecessary cyclical downturn in the United States (and elsewhere) in 1985.

Goldstein and Haynes fail to understand that, in a world where financial markets are integrated, only the strong-currency country (in 1984, the United States) has the option of expanding its money supply without throwing exchange rates further out of alignment, and threatening a breakdown of the international payments mechanism. Whence the need for explicit or implicit monetary coordination among the world's principal central banks.

## REFERENCES

- Goldstein, H. N. and Haynes, S. E., "A Critical Appraisal of McKinnon's World Money Supply Hypothesis," *American Economic Review*, March 1984, 74, 217–24.
- McKinnon, R. I., "Currency Substitution and Instability in the World Dollar Standard," *American Economic Review*, June 1982, 72,

320-33.

\_\_\_\_\_, *An International Standard for Monetary Stabilization*, Cambridge: MIT Press, 1984.

\_\_\_\_\_, Radcliffe, C., Tan, K-Y., Warga, A. D. and Willet, T. D., "International Influences on the U.S. Economy: Summary of an

Exchange," *American Economic Review*, December 1984, 74, 1132-34.

Tan, Kong-Yam, "Flexible Exchange Rates and Interdependence: Empirical Implications for U.S. Monetary Policy," unpublished doctoral dissertation, Stanford University, 1984.



# U.S. Monetary Policy and the Exchange Rate: Reply

By HENRY N. GOLDSTEIN AND STEPHEN E. HAYNES\*

In our earlier note (1984), we challenge the view of Ronald McKinnon (1982) that changes in rest-of-world money have had a dominant impact on U.S. inflation and output over the last decade because of indirect currency substitution (*ICS*). Steven Ambler and McKinnon (1985) disagree, modifying the *ICS* model by introducing the exchange rate as an additional indicator of foreign monetary policy. In this reply, we again conclude that no convincing evidence exists to support *ICS* for the United States over the past decade.

We make the following points to Ambler and McKinnon.

1. We agree that lagged growth rates in rest-of-world money tend to explain movements in the U.S. *WPI* better than they explain movements in the U.S. *GNP* deflator. But the *WPI*, which weights heavily the tradeable goods sector, is a less significant policy target than the *GNP* deflator, which represents the price of domestic output. And, as we have shown, the *GNP* deflator is much more sensitive to domestic money growth than to external money growth. Accordingly, we remain skeptical about McKinnon's recommendation that the U.S. authorities should give rest-of-world money a significant weight in formulating their monetary policy. Nor do we view the relatively high sensitivity of the *WPI* to changes in rest-of-world money as supporting the pervasiveness of *ICS* rather than simply the existence of traditional open economy interdependence. Although *ICS* links economies through capital arbitrage effects on interest rates and money demand,

the elasticities approach links economies directly through the tradeable goods sector.<sup>1</sup>

2. Ambler and McKinnon argue that our twelve-quarter moving averages used for the money variables tend to suppress the cyclical (one to two-year) movements with which McKinnon was concerned. However, as we stated earlier (p. 222), reestimation after expressing the money variables in distributed lags rather than moving averages still shows that rest-of-world money plays at best a minor role in influencing the U.S. *GNP* deflator. We conclude that the moving averages do not merely reflect secular or trend effects.

3. We agree that lagged exchange rates when added to our simple St. Louis model tend to influence the U.S. price level with the correct sign.<sup>2</sup> But the domestic price level and the exchange rate are necessarily linked whenever domestic residents consume foreign goods and domestic products compete against actual or potential imports. Given these obvious conditions, we fail to see how verification of a statistical linkage between the domestic price level and the exchange rate lends special support to McKinnon's view of the world as opposed to any other.

4. As noted by A. H. Meltzer (1984), McKinnon's policy recommendation regarding the exchange rate could lead to unfortunate

\*Department of Economics, University of Oregon, Eugene, OR 97403-1202. We thank Randy Eberts, Michael Hutchison, Joe Stone, and Bob Traa for helpful comments.

<sup>1</sup>Also see M. S. Wallace (1984) regarding the sensitivity of various U.S. price indexes to movements in U.S. (world) money.

<sup>2</sup>The coefficients on the lagged exchange rates are generally significant if the contemporaneous exchange rate is included in the equation, but insignificant if the contemporaneous exchange rate is excluded from the equation. Since the U.S. price level no doubt influences the exchange rate contemporaneously, estimates with the contemporaneous exchange rate on the right-hand side reflect simultaneous equation bias.

results in certain circumstances. During 1983-84, for example, the dollar appreciated against the deutsche mark even though economic growth in the United States was stronger than in West Germany. McKinnon's formula would presumably have favored faster monetary growth in the United States than actually prevailed, and slower monetary growth in West Germany. Given the evidence that domestic money affects domestic prices and activity more strongly than external money (see our earlier paper and F. Spinelli, 1983), this formula could well have tended to overheat the U.S. economy and retard desirable expansion in West Germany.

In summary, we continue to find no convincing evidence that the recent sharp cycles of inflation and deflation experienced by the United States since 1970 as well as the recent real growth with much reduced inflation can be explained exclusively, or even largely, by McKinnon's model of indirect currency substitution. We still view exogenous real shocks and swings in domestic monetary and fiscal policies as the major determinants of U.S. inflation and output.

## REFERENCES

- Ambler, S. and McKinnon, R., "U.S. Monetary Policy and the Exchange Rate: Comment," *American Economic Review*, June 1985, 75, 557-59.
- Goldstein, H. and Haynes, S. E., "A Critical Appraisal of McKinnon's World Supply Hypothesis," *American Economic Review*, March 1984, 74, 217-24.
- McKinnon, R., "Currency Substitution and Instability in the World Dollar Standard," *American Economic Review*, June 1982, 72, 320-33.
- Meltzer, A. H., "Cures that are Worse than the Disease," *London Financial Times*, August 22, 1984, p. 9.
- Spinelli, F., "Currency Substitution, Flexible Exchange Rates, and the Case for International Monetary Cooperation," *International Monetary Fund Staff Papers*, December 1983, 30, 755-83.
- Wallace, M. S., "World Money or Domestic Money: Which Predicts U.S. Inflation Best?," *Journal of International Money and Finance*, August 1984, 3, 241-44.

# The Money Supply Announcements Puzzle: Comment

By BARRY FALK AND PETER F. ORAZEM\*

In a recent paper in this *Review* (1983), Bradford Cornell presented a survey of existing literature on the empirical relationship between weekly money supply announcements made by the Federal Reserve and changes in the spot prices of several financial instruments at the time of the announcement. Cornell sought to unify and extend the work done in this area by estimating a number of relationships which bear directly on this issue. Among his main conclusions are that "asset markets are efficient with respect to money supply announcements" since "only the unexpected component of the announcement is correlated with price changes," and that the unexpected component of money supply announcements has "a highly significant positive correlation" with short-term interest rates, but only after the October 6, 1979 change in Fed policy (p. 651). Both of these conclusions are at variance with results reported in similar studies by Jacob Grossman (1981), V. Vance Roley (1982), and Thomas Ulrich and Paul Wachtel (1981). All three find that unanticipated announcements matter in periods before October 6, 1979, and Roley and Ulrich-Wachtel find that anticipated announcements matter in at least some of their regressions.

The main difference between Cornell's study and those mentioned above is that more exact measures of the change in short-term interest rates are used in the latter studies. Since the money supply announcements are made at 4:00 P.M., Cornell uses the change in the yield on three-month Treasury bills from 3:30 P.M. on the day of the announcement to 3:30 P.M. on the day after the announcement. Ulrich-Wachtel consider Treasury bill yield changes from 3:30 P.M. on the day of the announcement to 10:30 A.M. the next day. Grossman and Roley further refine the measure by looking at yield changes from 3:30 to 5:00 on the day of the announcement.

While the different measure of changes in the short-term interest rates might account for the inconsistencies between Cornell's study and earlier studies, the inconsistencies are still surprising. Undoubtedly, Cornell's measure of Treasury bill yield changes includes fluctuations in interest rates caused by new information on the day after the announcement. However, to the extent that markets function efficiently, only unexpected information should cause these additional interest rate movements. Because this unexpected information must be uncorrelated with past information, Cornell's estimates should be consistent and unbiased. Thus, it is puzzling that Cornell's results differ so drastically with the earlier studies.

This comment indicates that the inconsistencies are not attributable to differences in measures of short-term interest rate changes. Using essentially the same data and methods as Cornell, we obtain results that are more in line with the earlier studies. Below we briefly summarize the data and methodology and then contrast our results to those Cornell reported.

Like Cornell, we used the median value of the weekly Money Market Services survey for  $M1$  as our measure of the anticipated component of the announcement ( $EM_t$ ) and the difference between the actual announcement and its anticipated value as our measure of the unanticipated component of the announcement ( $UM_t$ ).<sup>1</sup> The immediate re-

<sup>1</sup>The Money Market Services data associate the anticipated announcement with the week to which the announcement actually pertains. Thus, for example, a money supply announcement made on Thursday, April 19, 1984, would be listed under "Statement Week April 9, 1984." We would look at the change in the Treasury bill rate from the close on the 19th to the close on the 20th as the response to this announcement. When the Fed's announcement occurs on a Friday, we consider the difference between the Friday close and the Monday close. More generally, we looked at the close on the first market-operating day after the announcement minus the close on the day of the announcement.

\*Assistant Professors of Economics, Iowa State University, Ames, IA 50011.

TABLE 1—SUMMARY STATISTICS

	Mean	Standard Deviation
Sample Period: January 5, 1978–October 4, 1979		
Cornell		
EM	0.12	0.45
UM	0.20	0.35
DTB	2.58	10.61
Falk-Orazem		
EM	0.21	0.35
UM	-0.12	0.44
DTB	2.81	10.48
Sample Period: October 11, 1979–December 18, 1981		
Cornell		
EM	0.06	0.26
UM	-0.04	0.58
DTB	8.68	40.80
Falk-Orazem		
EM	0.07	0.26
UM	0.04	0.54
DTB	6.46	40.66

Notes: UM = The unexpected component of the money supply announcement in percentage change in the money supply; EM = the anticipated component of the money supply announcement in percentage change in the money supply; DTB = the change from the close before the money supply announcement to the close after the announcement of the yield on the latest three-month Treasury bill in basis points.

sponse of the three-month Treasury-bill yield ( $DTB_t$ ) was measured as the difference in the 3:30 closing yield on the day of the announcement and the 3:30 close on the following (market) day. Summary statistics for our data and Cornell's (p. 650) are presented in Table 1.

If one reverses the line corresponding to anticipated and unanticipated money supply announcements for the first sample period in Cornell's summary, then our data on the money supply are virtually identical except for the sign differences on UM (which we computed by subtracting EM from the actual announcement). The standard deviations on the Treasury bill rate changes are nearly identical although our mean values differ somewhat. Since our data sources are identical and we thoroughly hand-checked our punched data against the original sources, we suspect that Cornell's sample mean for DTB during the second sample period should probably read "6," rather than "8."

To estimate the effect of the money supply announcement on the three-month Treasury

bill yield, Cornell regressed  $DTB_t$  on  $EM_t$ ,  $UM_t$ , and a constant for each of the two sample periods. We did likewise. The results from our regressions are compared to Cornell's (p. 651) in Table 2. Our fits (as measured by the  $R^2$ ) are substantially higher than those Cornell reports for both sample periods. We found that unanticipated money supply changes have a significant positive effect over both sample periods and anticipated money supply changes have a significant negative effect during the second sample period. The only significant coefficient Cornell found was the coefficient on unanticipated money over the second sample period.

Roley found that over the period September 9, 1977–October 4, 1979, unanticipated money supply changes had a significantly positive effect on Treasury bill yields while anticipated money supply changes had a negative, but not significant effect. This result persisted over the sample period October 11, 1979–January 31, 1980. Over the sample period February 8, 1980–November 20, 1981, Roley found both anticipated and unanticipated money supply announcements to have significant negative and positive effects, respectively.<sup>2</sup> Thus Roley's results look very much like our own.<sup>3</sup>

We are led to conclude that the results reported by Cornell regarding the effects of anticipated and unanticipated money supply

<sup>2</sup>Roley obtained the following results:

Sample period: September 29, 1977–October 4, 1979

$$DTB = -0.0027 + 0.0065 UM_t - 0.0014 EM_t, R^2 = .05$$

(0.0045) (0.0025) (0.0031)

Sample period: October 11, 1979–January 31, 1980

$$DTB = 0.0014 + 0.0510 UM_t - 0.0070 EM_t, R^2 = .34$$

(0.0205) (0.0161) (0.0223)

Sample period: February 8, 1980–November 20, 1981

$$DTB = 0.0160 + 0.0657 UM_t - 0.0531 EM_t, R^2 = .34$$

(0.0230) (0.0096) (0.0210)

(Standard errors are in parentheses.) Roley measures  $UM_t$  and  $EM_t$  in terms of the change in the money supply in billions of dollars.

<sup>3</sup>Estimates by Grossman and Urlich-Wachtel over the pre-October 6, 1979 sample period are also similar to ours, i.e., they obtain positive coefficients on unanticipated announcements and negative coefficients on anticipated announcements.

TABLE 2—REGRESSION COMPARISONS<sup>a</sup>  
 $DTB_t = a_0 + a_1 UM_t + a_2 EM_t + U_t$

	$a_0$	$a_1$	$a_2$	$R^2$
Sample period: January 5, 1978–October 4, 1979				
Cornell	2.30 (1.72)	1.92 (0.77)	0.24 (0.07)	.063
Falk-Orazem	4.66 (3.62)	7.09 (2.93)	-4.64 (-1.53)	.104
Sample period: October 11, 1979–December 18, 1981				
Cornell	7.21 (1.97)	30.46 (4.33)	-5.36 (-0.35)	.234
Falk-Orazem	6.75 (2.03)	41.83 (6.90)	-27.37 (-2.16)	.303

Notes: See Table 1.

<sup>a</sup>t-statistics are shown in parentheses.

announcements on the price of Treasury bill yields are incorrect quantitatively and qualitatively. Furthermore, we find that the measured response of short-term interest rates to money supply announcements are robust to slight changes in the measurement of the interest rate changes. Our results suggest that studies of money supply announcement effects are not as sensitive to the specification of Treasury bill rate changes as Cornell's results would suggest.

#### REFERENCES

- Cornell, Bradford, "The Money Supply Announcements Puzzle: Review and Interpretation," *American Economic Review*, September 1983, 73, 644–57.
- Grossman, Jacob, "The Rationality of Money Supply Expectations and the Short-Run Response of Interest Rates to Monetary Surprises," *Journal of Money, Credit and Banking*, November 1981, 13, 409–24.
- Roley, V. Vance, "The Response of Short-Term Interest Rates to Weekly Money Announcements," unpublished working paper, Federal Reserve Bank of Kansas City, 1982.
- Urich, Thomas and Wachtel, Paul, "Market Response to the Weekly Money Supply Announcements in the 1970s," *Journal of Finance*, December 1981, 36, 1063–72.
- Cornell, Bradford, "The Money Supply Announcements Puzzle: Review and Inter-

# The Money Supply Announcements Puzzle: Reply

By BRADFORD CORNELL\*

I thank Barry Falk and Peter Orazem for the comment on my 1983 paper. In response to their work I hand-checked my data using independent sources and found several errors. Initially I used a time-series of survey forecasts and forecast errors provided by Money Market Services. Though the survey data was correct, the reported forecast errors, in several instances, were based on incorrect numbers for the actual change in the money supply.

The main effect of these errors is to change the regression coefficients reported in Table 3 of my paper. For this reason a revised version of Table 3 (see Table 1) is presented here. The coefficients in the revised table are

very similar, but not identical, to those reported by Falk and Orazem. The remaining differences must be due to the fact that the interest data are not identical.

The only substantive change in the results is that expected money is now significant for long-term bonds in the post-October 6 sample, and nearly significant for short-term bills in both sample periods. This is more consistent with the findings of Falk and Orazem and the earlier work of Jacob Grossman (1981), V. Vance Roley (1982), and Thomas Ulrich and Paul Wachtel (1981).

Finally, Falk and Orazem are also correct in noting a typographical error in Table 2 of my earlier paper. The standard deviation of

TABLE 1—MONEY SUPPLY ANNOUNCEMENTS AND CHANGES IN ASSET PRICES<sup>a</sup>

$$DA_t = a_0 + a_1 UM_t + a_2 EM_t + U_t$$

<i>DA</i>	<i>a</i> <sub>0</sub>	<i>a</i> <sub>1</sub>	<i>a</i> <sub>2</sub>	<i>R</i> <sup>2</sup>
Sample Period: January 5, 1978–October 4, 1979				
<i>DTB</i>	4.31 (3.35)	6.69 (2.77)	-4.31 (-1.41)	.093
<i>DLTB</i>	1.03 (3.08)	1.41 (2.23)	-0.83 (-1.03)	.061
<i>DSP</i>	0.01 (0.42)	-0.28 (-1.81)	0.33 (1.71)	.061
<i>DDM</i>	-0.001 (-1.25)	0.14 (1.05)	0.26 (1.54)	.061
Sample Period: October 11, 1979–December 18, 1981				
<i>DTB</i>	5.65 (1.64)	38.48 (6.39)	-19.51 (-1.51)	.271
<i>DLTB</i>	4.54 (2.82)	14.25 (5.07)	-15.04 (-2.49)	.211
<i>DSP</i>	-0.001 (1.12)	-0.48 (-2.51)	0.16 (0.39)	.054
<i>DDM</i>	-0.005 (-0.58)	-.35 (-2.30)	0.52 (1.60)	.061

<sup>a</sup>Revised Table 3 (1983, p. 651); *t*-statistics are shown in parentheses.

\*Professor of Finance, Graduate School of Management, University of California, Los Angeles, CA 90024.

the change in the Treasury bill rate, *DTB*, for the post-October 6 period should be 6.68, not 8.68.

#### REFERENCES

- Cornell, Bradford, "The Money Supply Announcements Puzzle: Review and Interpretation," *American Economic Review*, September 1983, 73, 644-57.
- Grossman, Jacob, "The Rationality of Money Supply Expectations and the Short-Run Response of Interest Rates to Monetary Surprises," *Journal of Money, Credit, and Banking*, November 1981, 13, 409-24.
- Roley, Vance, V., "The Response of Short-term Interest Rates to Weekly Money Announcements," unpublished working paper, Federal Reserve Bank of Kansas City, 1982.
- Urich, Thomas J. and Wachtel, Paul, "Market Response to Weekly Money Supply Announcements in the 1970s," *Journal of Finance*, December 1981, 36, 1063-72.
- Falk, Barry and Orazem, Peter F., "The Money Supply Announcements Puzzle: Comment," *American Economic Review*, June 1985, 75, 562-64.

# Fisher's Paradox: Comment

By PATRICK HONOHAN\*

In a recent paper in this *Review* (1983), Jeffrey Carmichael and Peter Stebbing (C-S) examine the correlation between after-tax interest rates and the expected inflation rate. They contrast the Fisher hypothesis, that nominal interest rates move in line with expected inflation, with their own "inverted Fisher hypothesis," that nominal interest rates (on financial assets) are uncorrelated with expected inflation.

The main evidence adduced by C-S is a regression equation relating changes in interest rates to subsequent changes in the rate of inflation. Their equation can, with some manipulation, be rewritten in the simple form:

$$(1) \quad i_t - i_{t-1} = \alpha(\pi_t - \pi_{t-1}) + u_t,$$

where  $i_t$  is the nominal after-tax return on Treasury bills, and  $\pi_t$  is the rate of *CPI* inflation in the quarter after the observation of  $i_t$ . For both the United States and Australia, the coefficient  $\alpha$  is estimated not to be significantly different from zero. For instance, their equation 1.B (Table 1, p. 624) can be rewritten as

$$(2) \quad i_t - i_{t-1} = \frac{-0.024}{(T=1.1)} (\pi_t - \pi_{t-1})$$

(U.S., 1953:I–1978:IV,  $D-W=1.92$ .)

In these equations, the rate of inflation is thought of as a proxy for the expected rate of inflation. A natural reaction would be to conclude that the change in subsequent inflation has turned out to be a rather poor proxy for the change in expected inflation. However, C-S claim that the errors-in-variables bias in the estimate of  $\alpha$  is small, even under the Fisher hypothesis of  $\alpha=1$ .

In order to obtain an estimate of this bias, C-S assume that changes in inflation are truly a first-order moving average process, and that inflationary expectations are optimal forecasts relative to that process. Using conventional assumptions concerning the disturbance  $u_t$ , they arrive at a figure of  $-0.12$  for the bias in the estimate of  $\alpha$ . Unfortunately, this figure would only be correct for a regression based on the explanatory variable being in levels, whereas their equation in fact has first differences.<sup>1</sup>

The asymptotic bias for first differences is much larger. It is easy to see why: using actual inflation as a proxy for expected inflation involves an error whose variance is small relative to the variance of expected inflation. But the *change* in expected inflation is a rather smooth series by comparison with the change in actual inflation, so using the change in actual inflation as a proxy for the change in expected inflation involves an error whose variance is large relative to the variance of the change in expected inflation.

If we implement the general approach of C-S for the first difference of inflation, we find the much higher figure of about  $-0.85$  for the bias<sup>2</sup> (under the Fisher hypothesis). Of course, this would bring the point esti-

<sup>1</sup>Some other equations are in levels, but with correction for high positive first-order autocorrelation coefficients, leading to essentially the same problem with the calculation of bias as I outline here.

<sup>2</sup>Assuming that the first-order moving average parameter in the process generating expected inflation is about  $-0.75$ , and working in logs rather than levels as C-S do. There is little to choose between the logs and levels specification and only the log specification allows an analytic expression for the bias. In correspondence, Carmichael has stressed that my estimate of  $-0.85$  requires the usual assumption of no correlation between the (change in) true expected inflation series and the disturbance term. If this assumption is not valid, then not only might the bias be even greater, but the whole C-S approach, including their equation 1.G (Table 1) is marred by the inconsistency of *LS* estimators in the absence of this assumption.

\*Economic Adviser to the Taoiseach, Government Buildings, Dublin 2, Ireland. I am indebted to M. J. Harrison and L. O'Reilly for helpful comments.



mate of  $\alpha$  close to unity—the Fisher value.<sup>3</sup> Thus the equations reported by C-S could be read as broadly consistent with either the Fisher or the inverted Fisher hypothesis (if  $\alpha$  is truly zero, the bias is zero). Therefore we cannot take these equations as serious tests of either hypothesis.

Those who are not inclined to believe the inverted Fisher hypothesis may suspect that the bias is even greater than computed above. For one thing, working in first differences of inflation rates makes the choice of price index of primary importance. If the *CPI* were not exactly the index relevant to this Fisherian literature, and to the demand for Treasury bills, then its use as a proxy for inflationary expectations in this context would certainly be misleading. Changes in *CPI* inflation are only weakly (U.S.) or even

negatively (Australia) correlated with, for example, changes in producer price inflation. The above approach to the calculation of bias would not do where changes in the proxy variable were uncorrelated with the true but unobserved expectations variable, and in that case the expected value for the estimated coefficient  $\alpha$  would be zero.

One may also question the theoretical rationale given by C-S for their inverted Fisher theory. They argue convincingly that there is a high degree of substitutability between money and other financial assets. But they then infer that the implicit rate of return on holding money will be almost independent of the expected rate of inflation. This need only be so if the real stock of money were independent of the expected rate of inflation. This is not likely on theoretical or empirical grounds.

#### REFERENCE

<sup>3</sup>A referee has pointed out that in a market with rational inflation expectations, forecast errors will be serially uncorrelated. This would have the implication that, if expected inflation is highly positively autocorrelated, then the bias approaches  $-1$ .

Carmichael, Jeffrey and Stebbing, Peter W., "Fisher's Paradox and the Theory of Interest," *American Economic Review*, September 1983, 73, 619–30.

## Fisher's Paradox: Reply

By JEFFREY CARMICHAEL AND PETER W. STEBBING\*

In his comment on our 1983 paper on Fisher's paradox, Patrick Honohan argues that: 1) the use of first differences in our preferred equations leads us to seriously underestimate the potential errors-in-variables bias under Fisher's hypothesis we reported (fn. 9, p. 623); 2) the use of the *CPI* may make this potential bias even bigger; and 3) the implicit rate of return on holding money need not be independent of the expected rate of inflation.

The latter two points are quite trivial. With respect to the second point, all price series in Australia are highly correlated (contrary to Honohan's assertion)—this includes their rate of change and the change in their rate of change. Consistent with this observation, we tested several alternative price series and found the results to be invariant to the particular series chosen. We did not test alternative series for the United States. Given that all empirical tests (of which we are aware) of Fisher's hypothesis using U.S. data use the *CPI*, the onus would appear to be on Honohan to establish that the *CPI* is not a suitable index of prices, and that its use biases the results in a systematic way. His third point is equally irrelevant since the (approximate) independence of the implicit rate of return on money (our variable  $z$ ) and inflation is the hypothesis being tested. To assert that it may not be so, hardly advances our understanding of the issue at hand.

His first point, however, is more interesting. Our footnote 9 (p. 623) gives the general formula for the potential asymptotic errors-in-variables bias as  $b = -\sigma_\varepsilon^2 / (\sigma_{E\pi}^2 + \sigma_\varepsilon^2)$ . This formula is based on the assumption (our equation (6)) that observed inflation is the sum of expected inflation and a random error term:  $\pi_t = E\pi_t + \varepsilon_t$ .

Provided the process generating  $E\pi$  is independent of past forecasting errors, ( $\varepsilon_{t-i}$ ,  $i > 0$ ), the formula reported in footnote 9 is unaffected by first differencing the estimating equation or by correcting for serial correlation. If the process is not independent of  $\varepsilon_{t-i}$ , the calculation of the potential bias is affected. Since the calculations reported in footnote 9 use Douglas Pearce's ARIMA model for expected inflation, it is definitely incorrect.

We should emphasize at this point that it is not our estimates of the Fisher coefficient that are being questioned—it is our calculation of the potential bias in this coefficient that would exist if in fact Fisher were correct and we were wrong. To bring the issue into perspective we need to ask: first, how wrong might our calculation have been? And second, how important is the calculation to our conclusions?

The answer to the first is far from straightforward. Honohan suggests that using the general expression for the bias and Pearce's model for expected inflation gives a calculated bias closer to  $-0.85$  (our initial calculation was  $-0.12$ ). However, as he notes, this involves applying Pearce's coefficients to the log of prices rather than to the level. The need to do this arises because it is not possible to solve Pearce's actual model for a simple analytic expression for the bias. Applying Pearce's coefficients for the level of prices in a model for the rate of inflation involves a massive act of faith (indeed, we could not even go close to replicating Pearce's parameters when we tried to estimate an ARIMA model for the rate of inflation). Our hesitancy to embrace Honohan's calculations based on his arbitrary use of Pearce's coefficients reflects our belief in the critical importance of the model of expected inflation to the whole issue of measuring the potential errors-in-variables bias.

At the heart of the issue is the relative sizes of the variances of  $\varepsilon$  (the forecast error of the actual inflation rate) and  $E\pi$  (the time

\*International Department, Reserve Bank of Australia, Sydney 2001, Australia. The views expressed herein are our own, and are not necessarily shared by our employer.

path of expected inflation). If the process for  $E\pi$  depends *only* on past actual rates of inflation, the variance of  $E\pi$  will itself be a function of  $\epsilon$  only. For example, a simple  $AR(1)$  model,  $E\pi_t = \theta_0 + \theta_1\pi_{t-1}$ , gives  $\sigma_{E\pi}^2 = \theta_1^2\sigma_\epsilon^2/(1 - \theta_1^2)$ . In this simple model, the variance of expected inflation is infinite (and the potential bias very small) when  $\theta_1 = 1$  and zero (bias equal to minus one) when  $\theta_1 = 0$  (since  $\theta_1 = 0$  makes  $E\pi$  a constant).

It is a straightforward exercise to show that, with serial correlation in the estimating equation for the real interest rate, the size of the calculated potential errors-in-variables bias diminishes as other variables are added to the model for expected inflation. In the limit, expected inflation is independent of past prices and the expression in our footnote 6 (p. 633) is once again valid. All this emphasizes that one's estimate of the bias is critically dependant on the choice of model for expected inflation.

Finally, we turn to the importance of Honohan's point. So far we have emphasized the *potential* nature of the errors-in-variables bias. That is, it is only present if our null hypothesis of inverted Fisher is incorrect. In our equation 1.G (Table 1), we tested the alternative hypotheses using Pearce's expected inflation series directly. Since this equation does not contain a potential errors-in-variables problem, its support of the inverted Fisher hypothesis makes measurement of the potential bias irrelevant (or, at most, of second-order importance). The matter should rest there, but we are tempted to make one final point.

In our opinion, the errors-in-variables defense has been used as a crutch to explain why the data have consistently failed to support Fisher's hypothesis. From a purely technical perspective, those who believe in the Fisher effect for short-term financial interest rates can possibly find some solace in the errors-in-variables line of argument. Closer inspection of the data, however, should strain its credibility even to the most enthusiastic supporter. Our Figure 3 (p. 627), for example, tells a quite striking story. For Fisher to be correct, the true expected real after-tax rate of return would have to be a horizontal line through the middle of the graph, as would be the true expected rate of inflation. The difference between the true variables and the actual levels shown on the graph would need to be written off as erratic *ex post* movements, bearing no relationship to *ex ante* values. This might be convincing if those movements appeared to be erratic. In fact, what is striking about all three figures in our paper is the *systematic* nature of the movements. At least with respect to the data period in question, it is time for the crutch to be discarded.

## REFERENCES

- Carmichael, Jeffrey and Stebbing, Peter W., "Fisher's Paradox and the Theory of Interest," *American Economic Review*, September 1983, 73, 619-30.
- Honohan, Patrick, "Fisher's Paradox: Comment," *American Economic Review*, June 1985, 75, 567-68.

## Predictable Behavior: Comment

By RICHARD BOOKSTABER AND JOSEPH LANGSAM\*

Ronald Heiner's article in this *Review* (1983), argues for a new approach to the realm of behavior under uncertainty. He proposes that an action will only be added to the repertoire of actions if certain criteria of reliability are satisfied for that action. He argues that given increased uncertainty, the repertoire of actions will be limited. With few actions from which to choose, the individual's behavior will be more predictable. He then suggests that this explains behavior in a surprisingly wide range of applications, ranging from the evolution of social institutions to the structure of DNA.

In Heiner's enthusiasm, he has not thoroughly developed the analysis and given the helpful perspective that would come from relating his ideas to the expected utility-maximization framework that is already established in economic theory.<sup>1</sup> In particular, in his attempt for maximum abstraction to permit the most broad generalizations, Heiner gives no example of how his model could be used in extending the conventional framework of behavior under uncertainty. We will try to fill this gap by rephrasing his nomenclature, to the extent that we can properly interpret them, into a more classical framework.

### I. Reformulating the Analysis in an Expected Utility Framework

We will use a simple two-state model with a repertoire consisting of a single action,  $r$ , and with a known payoff matrix, giving values for taking the action  $r$  for states  $S=0$  and  $S=1$  of  $U(r,0)$  and  $U(r,1)$ , respec-

tively. The action,  $a$ , under consideration has the property that  $U(a,0) > U(r,0)$ , and  $U(a,1) < U(r,1)$ . The uncertainty induced by the environment is given by  $p(S=0)=p$  and  $p(S=1)=(1-p)$ . Perceptual uncertainty is introduced into the model by not permitting the individual to observe the state, but only to observe a random variable  $X$  with a distribution that depends upon the state. We let  $X$  take on the values 0 and 1. The probability that  $X=x$  given that  $S=s$  is given by  $p(X=x|S=s)$  or by  $p(x|s)$  when there is no chance for confusion. It should be noted that the fact that  $X$  is observed and that the states space  $S$  is a random variable incorporates the concepts of perceptual and environmental uncertainty.

Let us see how Heiner's model differs from this conventional model. Heiner begins by considering the "right" and "wrong" times to use an action. The right and wrong times depend upon the repertoire, which in this case has a single action. Using his notation, the right conditions for action  $a$  occur with probability  $\pi(e)$ . For action  $a$ , the right conditions occur when the state  $S$  is 0; therefore,  $\pi(e)=p$ .<sup>2</sup> In Heiner's paper, the term "right time" refers to uncertainty arising from the environmental variable only. The variable  $\pi(e)$  then cannot depend upon the distributional relationship between  $X$  and  $S$ . For an action  $a$  to be added to the repertoire, there must be a value  $x$  of the random variable  $X$  where the action  $a$  is preferred to the action  $r$ . We will therefore restrict our attention to the situation where the random variable  $X=x$ . In Heiner's framework, the right conditions for action  $a$  (i.e., when  $S=0$ ) are correctly recognized with probabil-

\*Morgan Stanley, 1251 Avenue of the Americas, New York, NY 10020, and Case Western Reserve University, Department of Mathematics and Statistics, Cleveland, OH 44106, respectively.

<sup>1</sup>Indeed, in Section I, Heiner suggests his analysis does not fit into the standard optimization paradigm.

<sup>2</sup>Heiner states: "Depending on the likelihood of different situations produced by the environment, the probabilities of the right or wrong time to select the action are written as  $\pi(e)$  and  $1-\pi(e)$ , respectively" (p. 565).

ity  $r(U)$ .<sup>3</sup> In our framework,  $r(U)$  equals  $p(X=x|S=0)$ . Similarly, the conditional probability of selecting the action  $a$  when it is actually the "wrong time,"  $w(U)$ , in our model is given by  $p(X=x|S=1)$ . The gains  $g(e)$  from selecting the action  $a$  under the right conditions are simply  $U(a,0)-U(r,0)$ , while the losses  $l(e)$  from selecting  $a$  under the wrong conditions are  $U(a,1)-U(r,1)$ . Note that when  $X=x$ ,  $\pi(e)r(U)$  is equivalent to  $p(X=x|S=0)=p(X=x, S=0)$ . Similarly,  $w(U)(1-\pi(e))=p(X=x, S=1)$ .

When will the action  $a$  be added to the repertoire of possible actions?—clearly when there is a value of  $X$  such that  $a$  would be chosen. Rephrasing his reliability condition, this will occur when for some  $x$ ,

$$\begin{aligned} & [U(a,0)-U(r,0)]p(S=0, X=x) \\ & + [U(a,1)-U(r,1)]p(S=1, X=x) \geq 0. \end{aligned}$$

If we divide by  $p(X=x)$  and rearrange terms, we find that for some  $x$ , it must be the case that

$$\begin{aligned} & U(a,0)p(X=x|S=0) \\ & + U(a,1)p(X=x|S=1) \\ & \geq U(r,0)p(S=0|X=x) \\ & + U(r,1)p(S=1|X=x). \end{aligned}$$

That is, an action  $a$  will be added for this repertoire only if there is a set of information for which it has higher expected utility than the action that is already in the repertoire.<sup>4</sup>

<sup>3</sup>Heiner states: "The conditional probability for selecting the action when it is actually the right time is written  $r(U)$ , where the likelihood of so doing depends on the structure of uncertainty,  $U=u(p, e)$ " (p. 565).

<sup>4</sup>While Heiner's concepts of  $g(e)$  and  $l(e)$  are too amorphous to model in the more general case of an arbitrary repertoire, it seems reasonable to assume that his intent is for an action  $a$  to be considered only if there is some value of  $X$  such that the expected utility of  $a$  is greater than the supremum of the expected utility of all other actions in the repertoire. Footnote 18 states performance "...may involve some kind of average over actions and/or environmental conditions" (p. 566), which certainly would suggest expected utility. He

This reliability condition appears to be identical to the condition that will arise out of a simple two-state expected utility-maximization model: an action will not be considered if there does not exist a situation where that action has expected utility that is at least as great as the other available actions.

## II. Comments on Model Specification

Heiner's model, recast in these terms, seems to fit quite comfortably in the conventional framework of expected utility maximization. In many important respects, it gives results that appear to be identical to those of the expected utility-maximization model. However, his approach does lead to some difficulties.

First, his "right time/wrong time" analysis can lead to strange results when his two-action model is extended to include a third action. In particular, in his model it may never be the right time to do action  $a_0$  even when the expected utility of this action dominates that of the other two actions.

If, for example, the repertoire consists of the actions  $q$  and  $r$ , where  $U(q,0) > 0 > U(r,0)$  while  $U(q,1) < 0 < U(r,1)$ , then it would seem that it would never be the right time for any action  $a$ , where  $U(a,0) < U(q,0)$  and  $U(a,1) < U(r,1)$ . If the probabilities that  $S=0$  given  $X=0$  and that  $S=0$  given  $X=1$  are strictly between zero and one, it is easy to construct choices for  $U(a,0)$ ,  $U(q,0)$ ,  $U(r,0)$ ,  $U(a,1)$ ,  $U(q,1)$ , and  $U(r,1)$ , such that  $U(q,0) > U(a,0) > 0 > U(r,0)$ ,  $U(q,1) < 0 < U(a,1) < U(r,1)$ , and where any reasonable observer would agree that action  $a$  dominates the actions  $q$  and  $r$ . The example can easily be constructed where, for either choice of  $X$ , the expected utility from choosing actions  $q$  and  $r$  is very negative while the expected utility when choosing action  $a$  is positive.<sup>5</sup>

also states that " $g(e)$  and  $l(e)$  still represent the gain or loss in performance from correct or mistaken selections, respectively...."

<sup>5</sup>It is possible that Heiner intended for his argument to be used not in determining when an action should be

This problem in Heiner's analysis results from having  $\pi(e)$  depend only upon the environmental variable. While  $e$  is the only argument of  $\pi(\cdot)$  in Heiner's paper, it is clear that this is a misspecification of the correct functional form; his variable  $p$  must also be included in  $\pi(\cdot)$ . To see this, note that the structure of uncertainty is represented by  $U = u(p, e)$ , and the conditional probabilities  $r(U)$  and  $w(U)$  are therefore also functions of both  $p$  and  $e$ . The unconditional probability  $\pi(\cdot)$  must therefore also have both  $p$  and  $e$  as arguments to have a functional form consistent with  $u(\cdot)$ .

But if  $\pi(e, p)$  is used in the reliability condition in place of  $\pi(e)$ , this will lead to problems of its own for Heiner's later analysis. The right-hand term of the reliability condition will no longer be constant as  $p$  varies, violating an assumption he uses throughout the many examples of the paper.<sup>6</sup> Specification of a tolerance limit (see his p. 566, and Figure 1) now becomes problematic, and it no longer becomes clear that "greater uncertainty [resulting from changes in  $p$ ] will in general require a more inflexible structure of rules" (p. 574), since both sides of reliability ratio may change with a change in  $p$ .<sup>7</sup>

---

added to the set of actions, but when a decision rule should be added to a repertoire of existing decision rules. In our illustration, a decision rule would be a function from the range of  $X$  into some specified set  $A$  of what we have called actions. Since the right time in Heiner's analysis is dependent only upon the uncertainty introduced by environmental variables, this would force the probability of it being the right time for a decision rule to be independent of the distributional relationship between the domain of the decision rule and the state of nature. This of course must lead to the same apparent contradictions as our earlier interpretation of Heiner's intent.

<sup>6</sup>It should be noted, however, that Heiner's fnn. 18 and 53, as well as the discussion in subsection E (p. 575), suggest an interaction between  $e$  and  $p$  that is not present in the analytical framework of the paper itself.

<sup>7</sup>There is also the issue of just what is meant by greater uncertainty. Nowhere in the paper is there presented a dynamic framework to show how the repertoire will change as uncertainty changes. His analysis as presented is for given levels of uncertainty, and does not contain a dynamic structure.

The rationale for using  $p$  and  $e$  rather than just one (noisy) information variable, such as  $X$  in the above example, is unclear. Both  $p$  and  $e$  are attacking the same factor, that of noise in the information. Either the environment is becoming less clear, so  $e$  is increasing, or the ability of the individual to extract information is diminishing, so  $p$  is decreasing. Either the world is growing darker or the individual's vision is dimming. Both situations lead to exactly the same result for the model: there is more noise in the information variable. This situation is easily treated in the more conventional model where the individual cannot observe  $p(S=s)$  directly, but can only observe  $p(s|x)$ . Adding noise simply requires a reevaluation of actions, as  $X$  becomes more noisy, there may be some actions that decline in expected utility for those values of  $X$  where they might have been taken, and they will then be dropped from consideration in the repertoire.

### III. Does Increased Uncertainty Necessarily Lead to a More Restricted Set of Actions?

While Heiner's principal thesis is that uncertainty leads to a reduced set of possible actions, it is actually equally plausible that some uncertainty will *increase* the total number of actions in the repertoire. It may be that with uncertainty, the actions currently in the repertoire remain optimal in some cases while other actions now become better in other cases. To illustrate this point, consider a single two-state, three-action example. Let  $U(a_1, s_1) > U(a_0, s_1) > U(a_2, s_1)$ , and  $U(a_2, s_2) > U(a_0, s_2) > U(a_1, s_2)$ . Thus,  $a_0$  is an "insured" action that leads to moderate utility in both states. Suppose  $X$  can take on three values,  $x_0$ ,  $x_1$ , and  $x_2$ , and further, that  $p(s_1|x_0) = p(s_1|x_1) = .99$ , and  $p(s_2|x_2) = .99$ . Here there is little uncertainty, and it is easy to find values of  $U(\cdot)$  which preserve the above inequalities and which will lead to the individual selecting action  $a_1$  when  $X = x_0$  or  $X = x_1$ , and selecting  $a_2$  when  $X = x_2$ . Now suppose we introduce more uncertainty by letting  $X$  be a weaker signal of the true state. If  $p(s_1|x_1) = .99$ ,  $p(s_2|x_2) = .99$ , while  $p(s_1|x_0) = .5$ , it is clear

that the values of  $U(\cdot)$  may be such that  $a_0$  will now be selected when  $X = x_0$ .<sup>8</sup> Thus, more uncertainty will lead to a richer set of actions.

#### IV. Does A Restricted Set of Actions Imply More Predictable Behavior?

While Heiner's model can be recast, and perhaps even benefited, in the conventional framework, he may consider the key implications he draws from the model to be even more important than the model itself. As his title suggests, his thesis is "that uncertainty becomes the basic source of predictable behavior." This thesis does not seem to follow from the implications of his analysis, however. The conclusion that he appears to try to draw in his analysis is that uncertainty leads to a restriction of the size of the repertoire of actions. It is not at all clear that choosing from a smaller set of actions leads to behavior that is more predictable. We would view predictability as coming from observations coinciding with the model in question. A restricted set of actions will not be predictable if the model of behavior suggests a large set of actions should be used.

Heiner fails to give a definition of predictability, and the concept of predictability that is central to his conclusions does not appear within his analytical framework. It is therefore difficult to address his claims regarding predictability directly. But as a simple illustration of the problem of making the leap from restricted behavior to predictable behavior, imagine an economist who models an agent's behavior as a function of some information variable  $X$ , and finds the optimal decision rule is to choose an action  $a$ , such that  $a = x$ ,  $x = 1, \dots, 10$ . If the individual follows the more restrictive set of actions,

$a = 3$  for  $x = 1, \dots, 6$  and  $a = 7$  for  $x = 7, \dots, 10$ , will his behavior be more predictable?<sup>9</sup>

These comments notwithstanding, we do not intend to discard the notion that uncertainty may lead to more inflexible or rule-bound behavior. Rather, we are suggesting some difficulties with Heiner's particular approach. Indeed, in our forthcoming paper, we have shown that a particular type of uncertainty, uncertainty which is not captured by the individual's model or world view, may lead to coarse decision rules, decision rules which do not make fine adjustments to information, and which therefore appear to be rigid or rule bound. With this uncertainty, an observer viewing the individual's behavior and comparing that behavior with the behavior which would be predicted by the individual's model will find unaccountable invariance to information: while the model indicates a fine adjustment to changes in information, the agent will be unaccountably coarse in his response. We agree with Heiner that such unexplained rigidity is a major problem in economic theory. And, as the theory becomes more sophisticated, the gap between the fine-tuned response that is predicted and the coarse response observed becomes more troublesome.<sup>10</sup>

#### V. Conclusion

Heiner's claims are bold and wide-ranging. But his model does not appear to differ from the classical optimization model in the way that leads to the conclusions of restricted variability of behavior he wishes to draw. Furthermore, it remains unclear whether restricted variability of actions will lead to the

<sup>8</sup>Intuitively, what is happening is that as the performance of the repertoire drops, that drops the standard of performance the individual compares new actions to, and lowers the tolerance limit. The reliability is lowered enough so a new action can be introduced the repertoire.

<sup>9</sup>On a more concrete level, consider schizophrenic behavior. Although typically characterized by a very limited set of actions, and actions that remain unchanged for wide variations in information, this behavior is extremely difficult to predict; it appears almost random in its relationship to the information received.

<sup>10</sup>This problem with complexity is mentioned early in his fn. 4 (p. 561).

greater predictability of behavior he implies. The true nature of his contribution might be easier understood if his analysis is placed in a more conventional model. In this way, it will also be easier to actually apply his claims to the existing models, and verify the validity of the many examples he presents in his paper. Our attempt in this comment has been to provide some aid in this regard.

## REFERENCES

- Bookstaber, Richard and Langsam, Joseph, "On the Optimality of Coarse Behavior Rules," *Journal of Theoretical Biology*, forthcoming.
- Heiner, Ronald A., "The Origin of Predictable Behavior," *American Economic Review*, September 1983, 73, 560-95.



# Predictable Behavior: Comment

By ROGER W. GARRISON\*

In a recent paper in this *Review* (1983), Ronald Heiner identifies "uncertainty in distinguishing preferred from less-preferred behavior" (p. 561) as the *sine qua non* of predictable behavior. Examples drawn from diverse disciplines (physical chemistry, biology, economics, and others) are evidence of the scope and richness of this insight. Yet, when the most basic law of economics, the Law of Demand, is reformulated in terms of Heiner's uncertainty-based behavior, the analysis becomes suspect. It fails to allow for the theoretical possibility of a Giffen good. The purpose of this comment is to argue that Heiner claims, at the same time, too much and too little for his new analytical framework. He claims too much by insisting on an unqualified inverse relationship between price and quantity demanded; he claims too little by failing to show how the qualifications needed in his own analytical framework parallel and complement the usual qualifications of standard price theory. Reconciling Heiner's Law of Demand with the more conventionally conceived Law of Demand can serve to defend standard price theory and to increase our confidence in the remainder of Heiner's analysis.

## I. Allowance for the Special Case of the Giffen Good

The Giffen good, whose existence in theory is rarely denied but whose existence in reality is in serious doubt, constitutes an exception to the general Law of Demand: an increase in price is accompanied by an increase, rather than a decrease, in the quantity demanded. This paradoxical price-quantity relationship is conventionally explained in terms of the relative strengths of the sub-

stitution effect and the income effect associated with a given price change.<sup>1</sup> If the income effect is opposite in sign and greater in magnitude in comparison to the substitution effect, price and quantity demanded will move in the same direction, and the good is a Giffen good. Acceptance of the conventional construction of demand curves, which accounts for both substitution effects and income effects, is an implicit admission of the possibility of a Giffen good—however unlikely its occurrence in reality. (By contrast, Heiner claims to have formulated the Law of Demand "without any qualification for income effects..." p. 580.)

To qualify the Law of Demand with an allowance for the special case of the Giffen good is to recognize the general limitations of the *ceteris paribus* assumption. If, in deriving the demand curve for a particular good, the *ceteris paribus* assumption is invoked in the strictest sense (which fixes income as well as all other prices, quantities, and expenditures), the resulting demand curve is a rectangular hyperbola.<sup>2</sup> That is, strict fixity in all other markets, including the "market" for cash balances, would leave as a residual a fixed expenditure in the market whose demand curve is being derived. To avoid this trivial result, numerous changes throughout other markets, each of negligible magnitude, are allowed for under the assumption of *ceteris paribus*. This permits the derivation of a nontrivial, downward-sloping demand curve.

The Giffen good is the result of circumstances under which the *ceteris paribus* assumption, even in the less-than-strict sense, cannot hold. For the income effect to be large enough to more than nullify the substitution effect, the good in question not only must be an inferior good but also must con-

\*Assistant Professor of Economics, Auburn University, Auburn, AL 36849. I am grateful to Don Bellante and Don Boudreaux of Auburn University for helpful suggestions on an earlier draft.

<sup>1</sup>See, for example, George Stigler (1966, p. 65).

<sup>2</sup>This problem with the *ceteris paribus* assumption is pointed out by Milton Friedman (1976, pp. 22–23).

stitute a substantial portion of the buyers' budgets. Grain or potatoes in underdeveloped economies are the most likely kind of goods to fulfill such specifications (see Stigler, 1947). Under these circumstances it is logically impossible to impound all other nonnegligible price and quantity changes in the *ceteris paribus* assumption. Such changes can be so substantial as to constitute a substantial change in real income whose effect, in theory, may outweigh the substitution effect. Alternatively stated, the consumers of, say, potatoes are so impoverished by a rise in the price of potatoes that they must even further restrict their consumption to this inferior good. As will be seen, this alternative statement allows us to see how the limitations of the *ceteris paribus* assumption implicit in Heiner's framework makes equal allowance for the existence, at least in theory, of the Giffen good.

## II. *Ceteris Paribus* in Heiner's Framework

In the analytical framework offered by Heiner, uncertainty in distinguishing between preferred and less-preferred behavior manifests itself as a gap between the "competence" of an economic agent and the "difficulty" in selecting the most preferred behavior (see his p. 562). The competence-difficulty gap, or *C-D* gap as Heiner calls it, forms the backdrop for the behavior of economic agents. Gains to behavior modification are attenuated by the probability that the environmental conditions which would justify such modifications exist and are correctly perceived; losses are similarly attenuated. Under conditions of uncertainty economic agents modify their behavior only when the prospective attenuated gains outweigh the prospective attenuated losses. Heiner summarizes the implicit decision rule with an inequality that relates the perception probabilities, which he calls the "reliability ratio," to the actual occurrence probabilities appropriately multiplied by the gain and loss functions, which he calls the "tolerance limit" (p. 556). Only when the reliability ratio exceeds the tolerance limit will economic agents actually modify their behavior.

The application of this decision rule to an environment in which prices are changing is

straightforward. A change in the price of a particular good will change the gain and loss functions associated with buying that good. Given the *C-D* gap, a fall in price will reverse the inequality for some economic agents causing additional agents to become buyers and existing buyers to buy additional quantities of the good. (Heiner's reasoning is slightly different from but compatible with my own: "a higher price requires purchasing behavior to be more reliable, which can be achieved only by reducing the probability of purchase" p. 580.)

It is important to recognize that in Heiner's framework the size of the *C-D* gap is impounded in a *ceteris paribus* assumption. Because it is difficult to even imagine a circumstance in which the competence of the agent and the difficulty in deciphering the environment move in the same direction and by the same "amount" (such that the *C-D* gap remains constant), the *ceteris paribus* assumption should be taken to mean that both *C* and *D* remain unchanged. In typical circumstances this assumption is fully warranted. There is no reason to believe that a change in a particular price has any discernible effect on either the complexity of the economic environment or the agent's competence to deal with it. Heiner's analysis is on as solid ground as standard price theory.

But paralleling the standard theory of demand, the Giffen good is associated with an atypical circumstance in which Heiner's *ceteris paribus* assumption cannot be maintained. When impoverished consumers are faced with an increased price in the one staple that makes up a large portion of their budgets, their choice set is affected in a non-negligible way. In the framework of standard price theory, we would say that their real incomes have been substantially reduced; in Heiner's framework, we would say that the complexity of the environment, and hence the difficulty in deciphering it, has been reduced. There is no reason to believe that the agents' "competence" has changed, but because of the new relationship between their incomes and the higher price of the staple good, their newly preferred behavior may become starkly clear: they must further limit their consumption to the staple good. Thus, the *C-D* gap, which must remain constant if

Heiner's derivation of the Law of Demand is to yield unambiguous results, has actually been reduced in a nonnegligible way as a result of the price increase. Although this reduction of the *C-D* gap does not, by itself, imply that agents will necessarily buy more as a result of a price increase, it does allow us to reconcile Heiner's analysis with the theoretical possibility of a Giffen good.

### III. Conclusion

Heiner makes the claim that in the simplicity of his logic is a "clear, unambiguous implication of the Law of Demand, which we have never been able to derive with traditional optimizing methods" (p. 580). But allowing for the possibility of a Giffen good does not constitute a shortcoming of traditional methods; this theoretical possibility, rather, must be allowed for in the traditional theoretical framework and in any other theoretical framework, including Heiner's. I have

shown that by focusing on the *ceteris paribus* assumption and its limitations, the Giffen good can be accommodated in Heiner's theory in a way that parallels the accommodation made in standard price theory. In both theoretical frameworks we can be sure that the Law of Demand holds only to the extent that the underlying *ceteris paribus* assumptions can be maintained.

### REFERENCES

- Friedman, Milton, *Price Theory*, Chicago: Aldine Publishing, 1976.
- Heiner, Ronald A., "The Origin of Predictable Behavior," *American Economic Review*, September 1983, 73, 560-95.
- Stigler, George J., "Notes of the History of the Giffen Paradox," *Journal of Political Economy*, April 1947, 55, 152-56.
- \_\_\_\_\_, *The Theory of Price*, 3rd ed., London: McMillan, 1966.

# Predictable Behavior: Reply

By RONALD A. HEINER\*

Richard Bookstaber and Joseph Langsam (1985), along with Roger Garrison (1985), have questioned certain parts of my paper (1983; hereafter called the "Origin" paper) wherein I suggested a new framework for explaining behavior. While my reply must inevitably focus on points of misunderstanding, nevertheless I would like to express appreciation for their comments. Moreover, errors on their part are partially due to the very brief development of formal theory in the paper.<sup>1</sup>

## I. Comparing Reliability Theory with the Conventional Framework

Bookstaber and Langsam begin by questioning whether reliability analysis extends the conventional decision theory framework. This issue is partially dealt with in my recent sequel to the Origin paper (1985c). I here focus on a key related topic not developed in that paper. To do so, think of the following conceptual scheme:  $A$  = the set of choosable actions, where individual actions may be randomized strategies over a set of more basic actions;  $B$  = the subset of  $A$  representing an agent's behavior;  $\gamma(Z)$  = the set of consequences resulting from actions  $Z \subset A$ , where individual consequences represent alternative probability distributions over a set of elementary events or outcomes;  $C^i(\gamma(Z))$  = the consequences chosen from  $\gamma(Z)$  according to a choice function  $C^i$ , where  $i$  refers to the type of choice function.  $C^i$  might be a standard expected utility function or a "nonexpected utility" function recently

developed by Machina, Chew, Fishburn, Loomes and Sugden, etc. (see Mark Machina, 1983).

Within this framework, the maximizing postulate is that agents act "as if" the consequences resulting from their behavior are selected from the set of choosable consequences  $\gamma(A)$  according to a choice function  $C^i$ ; that is,

$$(1) \quad \gamma(B) = C^i(\gamma(A)).$$

In other words, *theory about  $C^i$  is connected to observed behavior by postulating equality (1)*. But what happens if we discover behavior for which equality (1) fails to hold? This may already be the case for certain provisionally verified behavior patterns (such as *framing effects, probability ambiguity effects, preference reversals*, etc.; see Machina). How are we to deal with such possibilities? A typical response might be to introduce more subtle curvature assumptions or additional utility variables into a  $C^i$  function. Yet attempting to reestablish equality (1) in this way can do so only by steadily weakening the explanatory power of the theory. With this issue in mind, let us consider the *reliability condition* for an action  $a \in A$ ,

$$(2) \quad \frac{r_a}{w_a} > \frac{l_a}{g_a} \cdot \frac{1 - \pi_a}{\pi_a} = T_a,$$

where  $r_a = p(a|R_a)$ ,  $w_a = p(a|W_a)$ ,  $\pi_a = p(R_a)$ ,  $1 - \pi_a = p(W_a)$ ;  $R_a$  = the "right" conditions for which selecting action  $a$  will at least maintain or raise performance compared to selecting only other actions besides  $a$ ;  $W_a$  = the "wrong" conditions for which selecting action  $a$  will reduce performance compared to selecting only other actions besides  $a$ ;  $l_a$  = the "loss" in performance when action  $a$  is selected under the wrong conditions  $W_a$ ;  $g_a$  = the "gain" in performance when action  $a$  is selected under the right conditions  $R_a$ .

\*Member, 1984-85, The Institute for Advanced Study, Princeton NJ 08540; and Department of Economics, Brigham Young University.

<sup>1</sup>The Origin paper focused on broad conceptual issues and related applications, thereby shrinking mathematical details to a minimum. For further modeling and experimental analysis see my 1984, 1985a,b,c papers.

As discussed in the Origin paper, the ratio  $r_a/w_a$  measures agents' "reliability" in selecting action  $a$  (i.e., their ability to select action  $a$  under the right conditions  $R_a$  without mistakenly selecting it under the wrong conditions  $W_a$ ). The formula for  $T_a$  represents the minimum required reliability or "tolerance limit" that must be satisfied before agents can benefit from allowing themselves the option of choosing action  $a$  (i.e., from allowing themselves the flexibility to select action  $a$  along with other choosable actions).

Next consider how standard theory determines which actions to select. This is done by using an optimal decision rule,  $B^*: X \rightarrow A$ , that specifies how agents react to messages from a set of information  $X$  by choosing actions in their behavior repertoire  $A$ . Equality (1) implies  $B^*$  selects actions if and only if they maximize (according to a choice function  $C^i$ ) the "posterior" attainable performance contingent on received information. That is,  $B^*$  always selects an action  $a$  when messages are received for which it maximizes the posterior attainable performance contingent on receiving those messages (where such messages together represent the set  $R_a$  defined above); and conversely,  $B^*$  never selects action  $a$  when messages are received for which it does not maximize posterior attainable performance (where the latter messages represent the complement to  $R_a$  defined by  $W_a = X - R_a$ ). Thus,  $B^*$  or equality (1) implies  $r_a = 1$  and  $w_a = 0$  for all actions in  $A$ . Consequently, so long as equality (1) is assumed,  $r_a$  and  $w_a$  can only assume the values 1 and 0, respectively.

In contrast, reliability theory does *not* assume equality (1), but instead uses a choice function  $C^i$  to determine the size of the gain and loss variables ( $g_a$  and  $l_a$ ) of the tolerance limit  $T_a$  (see my 1984 paper for a precise discussion). Using a  $C^i$  function in this way implies no restriction on the  $r_a$  and  $w_a$  probabilities, thereby allowing one to investigate the larger set of possibilities defined by the inequality  $0 \leq w_a, r_a \leq 1$ . On the other hand, the above analysis implies equality (1) will in general be violated (except at the boundary of the preceding inequality). This conclusion can be formally stated by letting

$B^{Ci}$  denote the decision rule derived from a choice function  $C^i$  using the reliability condition. The following inequality can then arise,

$$(3) \quad \gamma(B^{Ci}) \neq C^i(\gamma(A)).$$

Now think of this in terms of the above mentioned experimental violations to equality (1). Inequality (3) means that the properties of choice functions like  $C^i$  do not necessarily carry over to the behavior  $B^{Ci}$  derived from them with the reliability condition (i.e.,  $C^i$  and  $B^{Ci}$  do *not* necessarily have the same properties). Consequently, *evidence that observed behavior violates equality (1) no longer contradicts using a  $C^i$  function to theoretically explain behavior.*<sup>2</sup> Instead, inequality (3) opens up the general possibility of modeling how and under what conditions the properties of  $B^{Ci}$  will deviate systematically from those satisfied by  $C^i$ . We can thereby retain the key analytical features of existing choice functions, yet use them to explain why observed behavior may fail to satisfy those same properties (see my 1984 paper).

The above discussion also applies to Bookstaber and Langsam's attempt to recast the reliability condition in terms of the conventional decision framework. They introduce so-called "perceptual uncertainty," but still in the traditional form represented by likelihood functions that give the probability of observing particular messages under different states of the environment. This leads them to argue that the  $p$  and  $e$  variables<sup>3</sup> introduced in the Origin paper are unnecessary for describing uncertainty. Standard theory suggests this view because (as dis-

<sup>2</sup>For example, observed violations of the Independence Axiom no longer contradict using an expected utility function to explain behavior.

<sup>3</sup>The  $p$  variables refer to agents' decision skills in reacting to information, and the  $e$  variables to the complexity-stability of their environment. Both sets of variables can be divided into subcategories. For example, the  $p$  variables include *perceptual skills* in observing and encoding potential information, *cognitive skills* in processing encoded information and deciphering how to respond to it, and *execution skills* in implementing behavior strategies suggested by the first two skills.

cussed above) it assumes agents optimally respond to information (i.e., standard theory models the perfect response to imperfect information).

However, no matter how rigorous and general such theory might appear, it only allows the  $r_a$  and  $w_a$  probabilities to be 1 and 0, respectively.<sup>4</sup> In contrast, the reliability condition enables one to investigate possibilities that extend well beyond the statistical properties of how information imperfectly reveals the environment. Consequently, giving up equality (1) or  $B^*$  does *not* mean we are left without a theory. Rather *it simply means we investigate the consequences of  $r_a$  and  $w_a$  away from the boundary values of 1 and 0*. Finally, note that we still have at our disposal major portions of existing statistical and utility theory to formally analyze how these probabilities change under different conditions. We can thereby model behavior without the maximizing postulate, yet still use the major analytical tools traditionally associated with it.<sup>5</sup>

## II. Uncertainty and Predictable Behavior

Let us next turn to another issue in Bookstaber and Langsam's comment about the appropriate meaning of predictable behavior. They suggest a predictability criterion which asks whether a behavior pattern is predicted correctly by an agent's theoretic-

cal model of what behavior (so that behavior mistakenly predicted by the model becomes *unpredictable* to the agent). By this criterion, someone with a faulty theory of planetary movement would regard their behavior as "unpredictable" even though their rigidly patterned celestial movements might be the very motivation for constructing the theory in the first place. As such, Bookstaber and Langsam's criterion confuses the difficulty of an observer detecting recurrent pattern in behavior with the difficulty of successfully explaining such pattern.

I do not wish to argue about what is the "correct" definition of predictability, except to say that I have the former idea (detecting recurrent pattern) in mind. This can be formalized by entropy measures used in statistical mechanics and information-cybernetics' theory. Let  $h_a$  denote the probability of selecting action  $a$ ,  $h_a = p(a)$ . Then agents' *behavioral entropy* equals the following sum over actions in  $A$ ,  $E^B = -\sum h_a \log h_a$ .  $E^B$  increases as the number of actions agents might select increases, but no single action is frequently chosen (reaching a maximum when all actions are equally likely to be selected; see Claude Shannon and Warren Weaver, 1963). Entropy measures have been extensively analyzed in several fields, thus enabling a number of key theorems and relationships to be readily accessed for new applications (for some examples, see my 1985c paper).

I now turn to the main thesis of the Origin paper whose validity is questioned by Bookstaber and Langsam; namely, the relationship between uncertainty and predictable behavior. To do so, let  $\rho$  denote the set of all  $r_a/w_a$  ratios for choosable actions in  $A$ . The set  $\rho$  describes an agent's reliability in reacting to information about when to select specific actions. The special case where agents are perfectly reliable is denoted  $\rho^\infty$ , meaning the  $r_a/w_a$  ratios are all infinite.  $\rho$  is said to be *bounded* if the reliability ratios of each action do not exceed a finite upper limit. In the case of  $\rho^\infty$ , agents still face "risk" due to imperfect information, but no additional uncertainty is involved. The term "uncertainty" thus refers to the case where agents' reliability  $\rho$  is bounded. In addition, let  $\rho^1$  denote

<sup>4</sup>One can also use likelihood functions of standard theory to measure the reliability of information used by agents (where the probabilities involved are unrestricted between 0 and 1). However, these probabilities do *not* measure agents' reliability in responding to information. See my 1985c paper for a two-stage reliability formula that combines both sources of uncertainty (from imperfect information and imperfect response to information).

<sup>5</sup>This also applies to using maximizing tools. For example, first-order conditions can be developed by differentiating  $r_a$  and  $w_a$ . However, doing so is redundant unless these probabilities can deviate from 1 and 0, which is possible only if decision rules besides  $B^*$  are used (otherwise the derivatives of  $r_a$  and  $w_a$  would always equal zero). In conceptual terms this means we can use  $r_a$  and  $w_a$  to derive optimizing conditions that are *conditional* on agents' true decision skills instead of having to abstract from them by assuming equality (1) or  $B^*$  (see my 1985a paper).

the other extreme case where agents are completely unable to distinguish right from wrong information for selecting different actions (so that they are equally likely to select actions under either  $W_a$  or  $R_a$ ). Thus,  $\rho^1$  means  $r_a/w_a=1$  for all  $a$ . The reliability sets describe a range of uncertainty possibilities, beginning with  $\rho^1$  at one extreme, and proceeding through intermediate cases where  $\rho$  is still bounded, finally limiting on  $\rho^\infty$  where only imperfect information remains.

Next consider the example constructed in Section III of Bookstaber and Langsam's comment. They introduce more risk (i.e., noisier information) in order to selectively reduce the performance achievable with two out of three possible actions, thereby expanding an agent's repertoire to include a third action. This is an instance of a more general possibility whereby two or more actions initially outperform another action for all potentially received information, but eventually fail to do so as the achievable performance from them drops sufficiently (due to worse information as in Bookstaber and Langsam's example, or because agents become less reliable at using information). Conversely, an action may initially dominate two or more other actions but fail to do so as agents have better information about when to select other actions or more reliably respond to such information.

It is possible that the transition from selecting a given action to selecting other ones may temporarily have agents benefiting from selecting all of them (i.e., their repertoire may initially expand to include a given action plus other ones before dropping out the former action; and vice versa). Consequently, such a transition may not always be monotonic. Note, however, the full transition necessarily involves a net shift between a given action and at least two other ones. This is because at least two other actions are needed to outperform a given action. If only one other action was involved, it would have to outperform the given action for all potentially received messages (thereby preventing that action from being selected in the first place).

The above considerations suggest that a pattern may emerge if we do not restrict

ourselves to special examples involving only small numbers of actions. Such is in fact the case. Two results can be proved with general assumptions about the relative numbers of different kinds of actions and how the performance resulting from them varies under different conditions.<sup>6</sup>

These results are briefly described as follows. Given agents' reliability in responding to information specified by  $\rho$ , individual actions are added to their behavioral repertoire only if the reliability condition for selecting them is satisfied. Conversely, actions can be deleted if at some point the reliability condition fails to hold. Additions and deletions continue until no excluded action satisfies the reliability condition and no included action violates it.<sup>7</sup> Finally, let  $E^B(\rho)$  denote the behavioral entropy resulting from this process for any given  $\rho$ . One can then prove two basic results:

A. Agents' behavioral entropy  $E^B(\rho)$  will drop from an initially positive level as  $\rho$  changes sufficiently toward  $\rho^1$ ; and conversely,  $E^B(\rho)$  will increase as  $\rho$  shifts sufficiently toward  $\rho^\infty$ .

B.  $E^B(\rho)$  is bounded if and only if  $\rho$  is bounded.

The results A and B describe a general pattern where sufficient changes in the reliability of responding to information will strictly raise or lower behavioral entropy away from its initial level. However, the shift between successively higher or lower entro-

<sup>6</sup>In particular, there must be enough variety of performance from different actions for there always to be a potential benefit from selecting more actions; there can only be a finite number of actions able to outperform another set of actions more often than a given positive fraction of the time, and the gains and losses from selecting individual actions cannot become arbitrarily large or small relative to each other.

<sup>7</sup>The order of adding or deleting particular actions need not be specified, or they can occur statistically (for example, the probability of adding or removing individual actions may depend on how much performance would change by doing so compared to that caused by adding or removing other actions). The theorem holds so long as the process cannot indefinitely add actions for which the resulting gain in performance is arbitrarily small compared to that of other actions which also satisfy the reliability condition but have not yet been included.

pies may not always be monotonic. Nevertheless, result B makes clear that is the boundedness of  $\rho$  (i.e., the presence of uncertainty) which limits agents' behavioral entropy. Thus, while the effects of uncertainty may not be continually monotonic, we can still regard uncertainty as the source of predictable behavior in the following sense:<sup>8</sup> without uncertainty affecting agents' decisions, recurrent pattern in their behavior (noticeable to an observer because  $E^B(\rho)$  is bounded) would not arise in the first place.<sup>9</sup>

Let us now briefly turn to two remaining points of lesser importance.

1) Bookstaber and Langsam argue that the reliability model is misspecified by not putting an agent's perceptual variables  $\mathbf{p}$  into the  $I_a, g_a, \pi_a$  components of the tolerance limit  $T_a$  (as first stated on p. 566 of the Origin paper). The possibility of  $\mathbf{p}$  variables affecting these components is, however, suggested later in the paper (see fnn. 18, 15; and p. 575). Noting the omission of  $\mathbf{p}$  from the initial specification is nevertheless a valid point. However, Bookstaber and Langsam also claim this omission produces substantive problems. This is mistaken. Explicitly putting  $\mathbf{p}$  into  $T_a$  simply makes clear that it may potentially affect both sides of the reliability

condition, as already the case for the  $\mathbf{e}$  variables.<sup>10</sup> (Note that this does not mean both sides of the reliability condition are noticeably or equally affected by every  $\mathbf{p}$  or  $\mathbf{e}$  variable.)

2) In Bookstaber and Langsam's Section II, an example with three actions is constructed; where there always exist states in which a first action is outperformed by one of the remaining two actions. This leads them to mistakenly interpret right and wrong circumstances ( $R_a$  and  $W_a$ ) to imply that only the remaining two actions can be chosen. However,  $R_a$  and  $W_a$  are *not* defined relative to the maximum potentially achievable performance from selecting other actions, but instead relative to that achievable given imperfect information and the reliability of responding to that information. Consequently, Bookstaber and Langsam's conclusion (that the reliability condition would not allow the first action to be chosen) is mistaken.

### III. The Law of Demand

Standard choice theory shows that an optimizing consumer can benefit from buying more when prices rise (called a Giffen response, or  $G$ -response) instead of always buying less (the latter response called the "Law of Demand," or  $L$ -response). However, the possibility of a  $G$ -response is usually ignored for practical purposes, largely because no credible evidence of Giffen behavior has even been documented (the Irish potato case notwithstanding; see Gerald Dwyer and Cotton Lindsay, 1984). Moreover, standard theory also suggests  $G$ -responses are unlikely since they can theoretically occur only for inferior goods that absorb more than a small fraction of consumers' total incomes. Nevertheless, a rea-

<sup>8</sup>Results A and B imply nonmonotonic effects of changing uncertainty are local deviations from a larger pattern that must nevertheless still hold. Consequently, instances of nonmonotonic effects like Bookstaber and Langsam's example do not contradict the basic relationship between uncertainty and predictable behavior. However, their example is still helpful in suggesting the need to qualify the intuitive language of the Origin paper. Accordingly, A and B specify the qualification while still affirming the paper's main thesis.

<sup>9</sup>The Origin paper also described the effects of uncertainty (see pp. 567–68) in terms of "behavioral rules" that restrict decision flexibility away from actions appropriate only for rare situations; or conversely, toward actions adapted to recurrent or "typical" conditions. An ambiguity may arise about the dual meaning of flexibility. Actions adapted to typical conditions can be described as "flexible" in handling a wide range of conditions (except unusual events that infrequently arise). In contrast, actions adapted only to infrequent conditions might be called "situation-specific" actions. Thus, behavioral inflexibility (*away* from situation-specific actions) also means agents flexibly adjust to (*only* the recurrent features of) their environment. The above was motivated by helpful discussions with Richard Langlois.

<sup>10</sup>As already mentioned above in Section I, Bookstaber and Langsam try to show that "noise in the information variable" can simulate the effects of both the  $\mathbf{p}$  and  $\mathbf{e}$  variables. However, if noisy information could really simulate  $\mathbf{e}$  (and  $\mathbf{e}$  can affect both sides of  $r_a/w_a > T_a$ ) then the components of  $T_a$  must thereby also depend on noisy information. This conclusion contradicts the opposite claim in Section I of their paper (that  $\pi_a$  cannot be affected by noisy information).



sonable interpretation of the theory would have to allow  $G$ -responses to be optimal at least some small fraction of the time (say, on the order of 3–5 percent of agents' reactions to different price changes over time).

Standard theory assumes agents are perfectly reliable at detecting when  $G$ -responses are more preferred than  $L$ -responses (corresponding to  $\rho^\infty$  as defined in Section II). Suppose, however, agents' decision skills are not sufficient to behave in this fashion. In particular, suppose agents are involved in an ongoing sequence of reactions to future price changes. A large number of interdependent factors affect the circumstances of individual price changes. For example, different commodities may be involved; spendable income or longer-term wealth may be different from one case to the next; the relative prices of other goods may be different; various psychological factors affecting their tastes may have changed, or ongoing experience with new consumption patterns may have altered previous tastes; the relative amounts spent on particular commodities may be different, including the one whose price changed; etc.

Note the last possibility. The fraction of income spent on particular goods may vary from case to case. Such variation is thus part of a large number of factors whose variation jointly contribute to the overall complexity of deciding how to respond to changing prices over time. Consequently, the difficulty of the decision problem is *not* defined assuming any given proportion of income spent on particular goods, *but instead already incorporates potential changes in this proportion as one of its many determining factors.*

Now compare this with Garrison's analysis. He assumes the fraction of income spent is like an exogenous parameter altering the difficulty of the decision problem from one case to the next, thereby violating implicit *ceteris paribus* assumptions of the Reliability Condition. This mistaken impression comes from viewing each price response in isolation (thus, permitting all sorts of changes in particular factors to possibly violate *ceteris paribus* assumptions). Once these isolated reactions are seen as part of a larger ongoing decision problem, the difficulty of such a dynamic problem must incorporate all

potential variations of decision factors that might appear in each specific instance.

This is the case even if a larger fraction of income spent increases the chance that  $G$ -responses are more preferred than  $L$ -responses (as possible for inferior goods). Such a criterion is not sufficient, but at best only necessary for  $G$ -responses to be more preferred. As already indicated, many other factors also affect when this is actually the case. Agents thus still face a nontrivial decision problem even for a subsequence of price reactions involving only goods that occupy sizable proportions of their future incomes.

I conclude with the following question: how can standard theory imply  $G$ -responses are optimal for some positive fraction of agents' ongoing price reactions, yet no stable frequency of Giffen behavior has been empirically discovered? Recall that  $h_G = p(G)$  is the chance of selecting a  $G$ -response, and  $\pi_G = p(R_G)$  is the chance of right conditions for  $G$ -responses (i.e., when they are actually more preferred than  $L$ -responses).

Standard theory implies agents select  $G$ -responses exactly as often as they are more preferred than  $L$ -responses, so that  $h_G/\pi_G = 1$  regardless of the size of  $\pi_G$ . On the other hand, when  $\rho$  is bounded, it can be shown for any action  $a$  that  $h_a/\pi_a \rightarrow 0$  as  $\pi_a \rightarrow 0$ . That is, the chance  $h_a$  of selecting an action must get arbitrarily small compared to the chance  $\pi_a$  of appropriate circumstances for selecting it as the latter probability gets smaller and smaller. Applying this to  $G$ -responses, we can easily have  $h_G$  vanishingly small compared to  $\pi_G$  just in the case where standard theory implies  $\pi_G$  is small but still positive. The latter result can be intuitively summarized as follows: the Law of Demand is a "behavioral rule" that would not have arisen if agents had the decision skills assumed in conventional choice theory (i.e., the extremely high predictability of the law owes its existence to uncertainty not present within this framework).

## REFERENCES

- Bookstaber, Richard and Langsam, Joseph, "Predictable Behavior: Comment," *American*

- Economic Review*, June 1985, 75, 571-75.
- Dwyer, Gerald P., Jr. and Lindsay, Cotton M., "Robert Giffen and the Irish Potato," *American Economic Review*, March 1984, 74, 188-92.
- Garrison, Roger W., "Predictable Behavior: Comment," *American Economic Review*, June 1985, 75, 576-78.
- Heiner, Ronald A., "The Origin of Predictable Behavior," *American Economic Review*, September 1983, 73, 560-95.
- , "On Reinterpreting the Foundations of Risk and Utility Theory," working paper, The Institute for Advanced Study, August 1984.
- , (1985a) "Uncertainty, Signal Detection Experiments, and Modeling Behavior," in Richard Langlois, ed., *The New Institutional Economics*, New York: Cambridge University Press, 1985.
- , (1985b) "Experimental Economics: Comment," *American Economic Review*, March 1985, 75, 260-63.
- , (1985c) "Origin of Predictable Behavior: Further Modeling and Applications," *American Economic Review Proceedings*, May 1985, 75, 391-96.
- Machina, Mark, "The Economic Theory of Individual Behavior toward Risk: Theory, Evidence and New Directions," Report No. 433, Center for Research on Organizational Efficiency, Stanford University, October 1983.
- Shannon, Claude E. and Weaver, Warren, *The Mathematical Theory of Communication*, Chicago: University of Illinois Press, 1963.

## Auditors' Report

February 22, 1985

Executive Committee  
The American Economic Association

We have examined the balance sheets of The American Economic Association as of December 31, 1984 and 1983, and the related statements of revenues and expenses, changes in general fund and restricted fund balances and changes in financial position for the years then ended. Our examinations were made in accordance with generally accepted auditing standards and, accordingly, included such tests of the accounting records and such other auditing procedures as we considered necessary in the circumstances.

In our opinion, the financial statements referred to above present fairly the financial position of The American Economic Association as of December 31, 1984 and 1983, its revenues and expenses and the changes in its financial position for the years then ended, in conformity with generally accepted accounting principles applied on a consistent basis.

Touche Ross and Co.  
Certified Public Accountants  
Nashville, Tennessee

## THE AMERICAN ECONOMIC ASSOCIATION BALANCE SHEETS, DECEMBER 31, 1984 AND 1983

	1984	1983
<b>Assets</b>		
CASH	\$ 915,479	\$ 705,732
INVESTMENTS, at market (Notes A and B)	3,165,188	3,227,487
ACCOUNTS RECEIVABLE, less allowance for doubtful accounts of \$1,854 (1984) and \$1,362 (1983)	101,250	174,787
INVENTORY OF <i>Index of Economic Articles</i> , at cost	90,918	47,992
PREPAID EXPENSES	22,383	21,287
OFFICE FURNITURE AND EQUIPMENT, at cost, less accumulated depreciation of \$25,688 (1984) and \$21,010 (1983)	40,676	43,634
	<u>\$4,335,894</u>	<u>\$4,220,919</u>
<b>Liabilities and Fund Balances</b>		
ACCOUNTS PAYABLE AND ACCRUED LIABILITIES	\$ 331,828	\$ 342,639
DEFERRED REVENUE (Note A):		
Life membership dues	47,058	49,680
Other membership dues	517,489	476,663
Subscriptions	417,104	401,254
<i>Job Openings for Economists</i>	18,138	17,004
	<u>999,789</u>	<u>944,601</u>
ACCRUAL FOR DIRECTORY (Note A)	211,610	147,141
FUND BALANCES:		
General	2,830,533	2,468,490
Unrecognized change in market value of investments (Notes A and C)	(147,997)	229,744
Net Worth	2,682,536	2,698,234
Restricted	110,131	88,304
Total Fund Balances	<u>2,792,667</u>	<u>2,786,538</u>
	<u>\$4,335,894</u>	<u>\$4,220,919</u>

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF REVENUES AND EXPENSES  
FOR THE YEARS ENDED DECEMBER 31, 1984 AND 1983

	1984	1983
<b>REVENUES FROM DUES AND ACTIVITIES:</b>		
Membership dues and subscriptions	\$ 781,598	\$ 757,999
Nonmember subscriptions	608,416	630,951
<i>Job Openings for Economists</i> subscriptions	28,240	27,247
Advertising	102,404	98,454
Sale of <i>Index of Economic Articles</i>	12,433	61,160
Sale of copies, republications, and handbooks	28,122	27,134
Sale of mailing list	38,989	34,517
Annual meeting	20,958	34,033
Sundry	52,847	51,365
	<u>1,674,007</u>	<u>1,722,860</u>
INVESTMENT GAINS (Note B)	<u>284,557</u>	<u>165,247</u>
<b>Net Revenues</b>	<b>1,958,564</b>	<b>1,888,107</b>
<b>PUBLICATION EXPENSES:</b>		
<i>American Economic Review</i>	475,735	480,228
<i>Journal of Economic Literature</i>	716,394	637,573
Directory publication (Note A)	65,000	60,000
<i>Job Openings for Economists</i>	49,859	49,754
<i>Index of Economic Articles</i>	9,854	32,958
	<u>1,316,842</u>	<u>1,260,513</u>
<b>OPERATING AND ADMINISTRATIVE EXPENSES:</b>		
General and administrative:		
Salaries	164,041	161,208
Rent	13,772	13,282
Other (Exhibit I)	162,197	156,401
Committee	53,711	44,585
Annual meeting	4,228	4,760
Provision for (benefit from)		
federal income taxes (Note A)	(3,000)	2,000
	<u>394,949</u>	<u>382,236</u>
<b>Total Expenses</b>	<b>1,711,791</b>	<b>1,642,749</b>
<b>REVENUES IN EXCESS OF EXPENSES</b>	<b>\$ 246,773</b>	<b>\$ 245,358</b>

See notes to financial statements.

## THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN GENERAL FUND BALANCE

	Total	Operations	Market Value Adjustments
<b>Balance at January 1, 1983</b>	<b>\$2,110,632</b>	<b>\$1,375,627</b>	<b>\$735,005</b>
Add market value adjustments resulting from inflation (Note A)	112,500	—	112,500
Add revenues in excess of expenses	245,358	245,358	—
<b>Balance at December 31, 1983</b>	<b>2,468,490</b>	<b>1,620,985</b>	<b>847,505</b>
Add market value adjustments resulting from inflation (Note A)	115,270	—	115,270
Add revenues in excess of expenses	246,773	246,773	—
<b>Balance at December 31, 1984</b>	<b><u>\$2,830,533</u></b>	<b><u>\$1,867,758</u></b>	<b><u>\$962,775</u></b>

See notes to financial statements.

## THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN RESTRICTED FUND BALANCE

	Balance at January 1	Receipts	Disburse- ments	Balance at December 31
<b>YEAR ENDED DECEMBER 31, 1983:</b>				
The Alfred P. Sloan Foundation and Federal Reserve System grants for increase of educational opportunities for minority students in economics	\$ 1,910	\$120,500	\$ 95,650	\$ 26,760
The Minority Scholarship Fund for minority students applying for graduate work in economics	5,000	—	—	5,000
The Rockefeller Foundation Grant for minority students applying for graduate work in economics	33,395	58,500	40,115	51,780
Sundry	2,331	5,300	2,867	4,764
	<b><u>\$42,636</u></b>	<b><u>\$184,300</u></b>	<b><u>\$138,632</u></b>	<b><u>\$ 88,304</u></b>
<b>YEAR ENDED DECEMBER 31, 1984:</b>				
The Alfred P. Sloan Foundation and Federal Reserve System grants for increase of educational opportunities for minority students in economics	\$26,760	\$141,570	\$ 98,690	\$ 69,640
The Minority Scholarship Fund for minority students applying for graduate work in economics	5,000	—	—	5,000
The Rockefeller Foundation Grant for minority students applying for graduate work in economics	51,780	890	22,680	29,990
Sundry	4,764	3,070	2,333	5,501
	<b><u>\$88,304</u></b>	<b><u>\$145,530</u></b>	<b><u>\$123,703</u></b>	<b><u>\$110,131</u></b>

See notes to financial statements.

THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF CHANGES IN FINANCIAL POSITION  
FOR THE YEARS ENDED DECEMBER 31, 1984 AND 1983

	1984	1983
Cash, beginning of year	\$705,732	\$640,267
SOURCES OF CASH:		
Revenues in excess of expenses	246,773	245,358
Noncash charges:		
Depreciation	3,780	3,677
Directory publication (Note A)	65,000	60,000
Market value adjustments (Note A)	<u>(20,262)</u>	<u>57,572</u>
Cash provided by operations	<u>295,291</u>	<u>366,607</u>
INCREASE (DECREASE) IN CASH DUE TO CHANGES IN:		
Investments	62,299	(419,402)
Accounts receivable	73,537	(98,412)
Inventory of <i>Index of Economic Articles</i>	(42,926)	(10,481)
Prepaid expenses	(1,096)	(6,265)
Office furniture and equipment	(822)	(20,035)
Accounts payable and accrued liabilities	(10,811)	(19,313)
Deferred revenue	55,188	(19,912)
Accrual for directory	(531)	-
Restricted funds	21,827	45,668
General fund, market value adjustments	115,270	112,500
Unrecognized change in market value of investments	<u>(357,479)</u>	<u>134,510</u>
Cash, end of year	<u>\$915,479</u>	<u>\$705,732</u>

See notes to financial statements

### Notes to Financial Statements

#### A. Summary of Significant Accounting Policies

*Investments* are accounted for on a market value basis. According to the method the Association uses to value investments, the change in market value of corporate stocks, government obligations, bonds and commercial paper during the year, after adjusting for an inflation factor (3.7% in 1984 and 4.2% in 1983), is recognized in income over a three-year period for corporate stocks and reflected in current income for government obligations, bonds and commercial paper.

*The Accrual for directory* results because every three to five years the Association publishes a directory which lists, among other things, the names and addresses of its membership. This directory was most recently published in 1981 and distributed at no cost to the membership. In order to properly match the publishing cost of this directory with revenue from membership dues, the Association provided \$65,000 in 1984 and \$60,000 in 1983 for estimated publishing costs which will reduce actual directory expenses in the year of publication.

*Deferred revenue* represents income from membership dues for and subscriptions to the various periodicals of the Association which are deferred when received. These amounts are then recognized as income following the distribution of the specified publications to the members and subscribers of the Association. Income from life membership dues is recognized over the estimated average life of these members.

*The American Economic Association* files its federal income tax return as an educational organization, substantially exempt from income tax under Section 501(c) (3) of the Internal Revenue Code. As required by Section 511(a) of this Code, the Association provides for federal income taxes on certain revenues which are not substantially related to its tax exempt purpose. This "unrelated business income" includes income from advertising and the sale of mailing lists. The Association has been determined to be an organization which is not a private foundation.

**B. Investments and Investment Income**

Investments consist of:

	December 31, 1984		December 31, 1983	
	Cost	Market	Cost	Market
Government obligations, bonds and commercial paper	\$ 923,345	\$ 983,784	\$ 847,212	\$ 908,212
Corporate stocks and mutual funds	1,696,178	2,181,404	1,649,062	2,319,275
	<u>\$2,619,523</u>	<u>\$3,165,188</u>	<u>\$2,496,274</u>	<u>\$3,227,487</u>

Investment gains recognized consist of:

	Year Ended December 31	
	1984	1983
Government obligations, bonds, and commercial paper:		
Interest	\$176,304	\$151,144
Increase (decrease) in market value recognized	(28,084)	(76,584)
	<u>148,220</u>	<u>74,560</u>
Corporate stocks and mutual funds:		
Cash dividends	87,992	71,675
Increase in market value recognized (Note C)	48,345	19,012
	<u>136,337</u>	<u>90,687</u>
Investment gains, net	<u>\$284,557</u>	<u>\$165,247</u>

**C. Unrecognized Change in Market Value of Investments**

As described more fully in Note A, the Association recognizes in income over a three-year period changes in the market value of its corporate stocks. The following summarizes the years in which market value changes in stocks occurred that affected 1984 and 1983 revenues, and the amount of these market value increases (decreases) that will be recognized in income in future periods.

Year of Market Value Change	Recognized in Income in		To be Recognized in		Unrecognized Change	
	1984	1983	1985	1986	1984	1983
1981	\$ -	(\$139,131)	\$ -	\$ -	\$ -	\$ -
1982	88,397	88,396	-	-	-	88,397
1983	69,747	69,747	71,600	-	71,600	141,347
1984	(109,799)	-	(109,799)	(109,798)	(219,597)	-
	<u>\$ 48,345</u>	<u>\$ 19,012</u>	<u>(\$ 38,199)</u>	<u>(\$109,798)</u>	<u>(\$147,997)</u>	<u>\$229,744</u>

The Association's revenues in excess of expenses would have been (\$130,968) in 1984 and \$437,440 in 1983 if changes in the market value of corporate stocks, after adjustment for inflation, had been recognized only, but entirely, in the year in which they occurred.

The Association's investment gains would have been \$121,978 in 1984 and \$128,168 in 1983 if 4% of the market value of the portfolio had been considered investment income (without regard to dividends, interest, capital gains and inflation). Under this method, the Association's revenues in excess of expenses would have been \$84,195 in 1984 and \$208,279 in 1983.



**D. Retirement Annuity Plan**

Employees of the Association are eligible for participation in a contributory retirement annuity plan. Payments by the Association and participating employees are based on the employee's compensation. Benefit payments are based on the amounts accumulated from such contributions. The total pension expense was approximately \$27,000 and \$23,000 for 1984 and 1983, respectively.

**E. Ratio of Net Worth to Expenses**

The ratio of net worth at December 31, 1984 to 1985 budgeted expenses is 1.43 and the ratio of net worth at December 31, 1983 to actual 1984 expenses is 1.58.

**EXHIBIT 1—THE AMERICAN ECONOMIC ASSOCIATION STATEMENTS OF OTHER  
GENERAL AND ADMINISTRATIVE EXPENSES FOR THE YEARS ENDED  
DECEMBER 31, 1984 AND 1983**

	1984	1983
Dues and subscriptions	\$ 49,600	\$ 39,925
Mailing list file maintenance	23,617	24,385
Postage	16,478	15,788
Periodic mailing expenses	14,189	14,118
Accounting and legal	12,140	12,750
Investment counsel and custodian fees	11,960	11,267
Office supplies	11,988	10,582
Insurance and miscellaneous	4,088	9,006
President and president-elect expenses	5,648	5,315
Telephone	4,626	5,033
Depreciation (straight-line method)	3,780	3,677
Currency exchange charges	1,548	2,160
Uncollectible receivables	1,002	1,658
Travel and entertainment	1,533	737
	<u>\$162,197</u>	<u>\$156,401</u>

## NOTES

### *1986 Nominating Committee of AEA*

In accordance with Section IV, paragraph 2, of the bylaws of the American Economic Association as amended in 1972, President Elect Alice M. Rivlin has appointed a Nominating Committee for 1986 consisting of W. Arthur Lewis, Chair; William J. Beeman, William J. Boyes, Padma Desai, Thomas J. Finn, A. Nicholas Filippello, and Marjorie Honig.

Attention of members is called to the part of the bylaw reading, "In addition to appointees chosen by the President-Elect, the Committee shall include any other member of the Association nominated by petition including signatures and addresses of not less than 2 percent of the members of the Association delivered to the Secretary before December 1. No member of the Association may validly petition for more than one nominee for the Committee. The names of the Committee shall be announced to the membership immediately following its appointment and the membership invited to suggest nominees for the various officers to the Committee."

### *Nominations for AEA Officers: 1986*

The Electoral College on March 22 chose Gary S. Becker as nominee for President Elect of the American Economic Association in the balloting to be held in the autumn of 1985. Other nominees (chosen by the 1985 Nominating Committee) are: Vice President (two to be elected), Richard N. Cooper, Peter A. Diamond, Mancur Olson, and Thomas C. Schelling; for members of the Executive Committee (two to be elected), Donald F. Gordon, Sherwin Rosen, Thomas J. Sargent, and T. N. Srinivasan.

Under a change in the bylaws as described in the *American Economic Review Proceedings*, May 1971, page 472, additional candidates may be nominated by petition, delivered to the Secretary by August 1, including signatures and addresses of not less than 6 percent of the membership of the Association for the office of President-Elect, and not less than 4 percent for each of the other offices. For the purpose of circulating petitions, address labels will be made available by the Secretary at cost.

Donation available: Full years of the *American Economic Review* and the *Journal of Economic Literature* (1969-74). Contact Dr. Sophie Korczyk, 706 Little Street, Alexandria, VA 22301.

The eighth annual Middlebury College Conference on Economic Issues, "The Psychological Foundations of Economic Behavior," will be held October 25-26, 1985. For further information, contact Paul J. Albanese, Pro-

fessor of Economics, Middlebury College, Middlebury, VT 05753 (telephone 802+388-3711, Ext. 5322).

The National Institute on Aging invites grant applications for research and research training that focus on increases in longevity at the later ages, and the future explosive growth of that segment of the population. Application deadlines are March 1, July 1, and November 1. Contact Ms. Mildred D. Mader, National Institute on Aging, Bldg 31C, Rm 4C32, 9000 Rockville Pike, Bethesda, MD 20205.

The University of Venice will host a conference on Advances in the Analysis of Economic Dynamic Systems in January 1986. The focus will be on applications to economics of recent developments in the theory of nonlinear dynamic systems, to include nonlinear oscillations, bifurcation theory, catastrophe theory, synergetics, and irregular oscillations, or chaotic dynamics. Emphasis will be on use of deterministic models to describe certain irregularities of real economies that have so far been explained by means of stochastic models. Prospective contributors should submit full texts of approximately 15 typewritten pages by September 1985. No expenses will be paid; however, there will be prizes of one million lira each for the two best papers. Contact the Conference Secretary: Marji Lines, Dipartimento di Scienze Economiche, Università di Venezia, Dorsoduro 3246, Venezia 30123.

The Inter-University Consortium for Political and Social Research will hold its twenty-third Summer Program in Quantitative Methods of Social Research in Ann Arbor, July 1-August 23, 1985. The Program is divided into two 4-week sessions: July 1-26 and July 29-August 23. Individuals can attend either or both, or can participate in workshops lasting one week or less. For full information and the complete list of Course Offerings/Workshops, contact ICPSR official representatives at member colleges and universities, or the Director, Henry Heitowit, Educational Resources, ICPSR Summer Program, P.O. Box 1248, Ann Arbor, MI 48106 (telephone 313+746-8392).

Harvard Law School offers fellowships to college and university teachers in the social sciences and humanities to enable them to study fundamental techniques, concepts, and aims of law, so that, in their teaching and research, they will be better able to use legal materials

and legal insights which are relevant to their own disciplines. Further information may be obtained from the Chairman, Committee on Liberal Arts Fellowships in Law, Harvard Law School, Cambridge, MA 02138.

---

*Call for Papers:* The annual meeting of the Association of Environmental and Resource Economists (AERE) will be held jointly with the AEA in New York City, December 28–30, 1985. Those interested in having papers considered should send two copies of a one-page abstract to the President of AERE, Professor V. Kerry Smith, Department of Economics, Box 52 B, Vanderbilt University, Nashville, TN 37235.

---

*Call for Papers:* The AEA Committee on the Status of Women in the Economic Professions (CSWEP) will sponsor two sessions at the 1985 Southern Economic Association meetings, November 24–26, 1985, in Dallas at the Hyatt Regency Hotel. Those wishing to present theoretical or empirical papers for either session: "Comparable Worth: Applied Microeconomic Theory," or "Gender Effects of Taxes, Pensions, and Fringe Benefits," should contact Professor Marie Lobue, Department of Economics and Finance, University of New Orleans, New Orleans, LA 70148 (telephone 504 + 286-6485).

---

*Call for Papers:* UNEP (United Nations Environment Programme), the Autonomous University of Nuevo Leon of Monterrey, Mexico and the University of Cincinnati are organizing an international Conference on the Economics of Environmental Protection in Mexico to be held in Monterrey, October 14–18, 1986. Papers on the economic aspects of environmental protection in developing countries in general and in Mexico in particular are invited. Submit a one-page abstract as soon as possible. Authors from Mexico and Latin America should contact Lic. Manuel Silos M., Director, Facultad de Economía, Universidad Autónoma de Nuevo Leon, Apartado Postal 288, Monterrey, N.L. 64000, Mexico; all others contact Professor Haynes C. Goddard, Department of Economics (371), University of Cincinnati, Cincinnati, Ohio 45213.

---

*Call for Papers:* The Fourth International Congress of the North American Economics and Finance Association will be held July 23–26, 1986, at the Université de Montréal. Economic and financial topics relevant to Canada, Mexico, the Caribbean, and the United States will be covered and will be held in English, Spanish, and French. Deadline for submission of papers is January 1, 1986. To present papers, be a discussant, or chair a session, contact M. C. Thiron, Chairman of the Program Committee, c/o G. Gaudichon, CRDE-Department of Economics, Université de Montréal, C.P. 6128,

Succ. "A", Montréal, Quebec H3C 3J7 (telephone R. Tremblay: 514 + 343-6549; or 514 + 343-6657).

---

The Council for International Exchange of Scholars (CIES) announces the 1986–87 competition for Fulbright Scholar Awards in research and lecturing abroad. Benefits include round-trip travel for the grantee and, for full academic year awards, one dependent; living costs allowance; tuition as well as book and baggage allowances. Grants are for 3–12 months in over 100 countries. Applicants must be U.S. citizens, hold the Ph.D. or comparable professional qualifications, and have university or college teaching experience. Deadlines are June 15, 1985 for Australasia, India, Latin American and the Caribbean; September 15, 1985 for Africa, Asia, Europe, and the Middle East; November 1, 1985 for Junior Lectureships to France, Germany, Italy, and Spain; December 1, 1985 for Administrators' Awards in Germany, Japan, and the United Kingdom; December 31, 1985 for NATO Research Fellowships; and February 1, 1986 for Seminar in German Civilization Awards, Spain Research Fellowships, and France and Germany Travel-Only Awards. For more information and applications, contact CIES, Eleven Dupont Circle, NW., Washington, D.C. 20036-1257 (telephone 202 + 939-5401).

---

The Indo-U.S. Subcommittee on Education and Culture is offering twelve long-term (6–10 months) and nine short-term (2–3 months) awards for 1986–87 research in India. Applicants must be U.S. citizens at the postdoctoral or equivalent professional level. Fellowship terms include \$1,500 per month, of which \$350 per month is payable in dollars and the balance in rupees; an allowance for books and study/travel in India; and international travel for the grantee. In addition, long-term Fellows receive international travel for dependents; a dependent allowance of \$100–\$250 per month in rupees; and a supplementary research allowance up to 34,000 rupees. The application deadline is June 15, 1985. For application forms and further information contact Council for International Exchange of Scholars (CIES), Attn: Indo-American Fellowship Program, Eleven Dupont Circle, Suite 300, Washington, D.C. 20036-1257 (telephone 202 + 939-5472).

---

The fourth annual International Conference on Women and Organizations will be held in San Diego, CA, August 8–9, 1985. The theme is "Promoting Career Progress for Women: Effective Programs for Individuals and Organizations." Practicing managers, consultants, and academics are invited to attend. For further information, contact Jean Ramsey, Department of Management, Western Michigan University, Kalamazoo, MI 49008 (telephone 616 + 383-1357).

*Call for Papers:* The thirteenth annual meeting of the Midsouth Academy of Economics and Finance will be held February 13–15, 1986, in Memphis Tennessee. Paper proposals should be submitted by October 15, 1985. Those interested in chairing, discussing, or organizing special sessions should also contact David E. R. Gay, President-Elect, Midsouth Academy of Economics and Finance, Department of Economics, BA402, University of Arkansas, AR 72701.

The Society for Risk Analysis will hold its annual meeting on the theme, "Enhancing Risk Management," to be held in Washington, D.C., October 7–9, 1985. For further information, contact Janice Longstreth, ICF, 1850 K Street, N.W., Washington, D.C. 20031 (telephone 202+862-7915).

Piero Sraffa has named Pierangelo Garegnani literary executor in his will, leaving him the rights on all his writings. Garegnani has initiated work for an edition of Sraffa's writings to include correspondence. Any material (photocopies, information, personal recollections) would be appreciated. Contact Professor Pierangelo Garegnani, Department of Public Economics, Faculty of Economics, University of Rome, Via del Castro Laurenziano 9, 00161 Rome, Italy.

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to airfare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States. To be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for application to be received in the ACLS office are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Attention: Travel Grant Program), 228 East 45 Street, New York, NY 10017, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting.

### Deaths

Paul W. Conner, InterFuture, New York, NY, November 28, 1984.

George Montgomery, Kansas State University (retired), October 26, 1984.

Emanuel Stein, professor emeritus, department of economics, New York University, January 26, 1985.

### Retirement

Carl Kreider, professor of economics, Goshen College, June 30, 1985.

### Foreign Scholars

Luis F. Gonzales Vigil, Universidad Nacional Mayor de San Marcos, Lima, Peru: visiting professor of Latin American Studies, University of Pittsburgh, winter 1985.

Yasukichi Yasauba, Osaka University: visiting professor, University of Pittsburgh, winter 1985.

### Promotions

Annette E. Meyer: associate professor of economics, Trenton State College, September 1984.

William E. Mitchell: professor of economics, University of Missouri-St. Louis, September 1, 1983.

Donald L. Phares: professor of economics and public policy administration, University of Missouri-St. Louis, September 1, 1984.

Sherrill Shaffer: senior economist, Federal Reserve Bank of New York, November 15, 1984.

### Administrative Appointments

Stephen Dresch, International Institute for Applied Systems Analysis, Austria: dean, School of Business and Engineering Administration, Michigan Technological University, January 1, 1985.

Annette E. Meyer: chair, department of economics, Trenton State College, January 1985.

Robert L. Sorensen: chairman, department of economics, University of Missouri-St. Louis, September 1, 1984.

### Appointments

Carol W. Jones, Harvard University: assistant professor of economics and natural resources, University of Michigan, September 1, 1984.

Hassan Khademian, Michigan State University: assistant professor of economics, University of Missouri-St. Louis, September 1, 1984.

Hugh H. Macaulay, Jr., National Taiwan University: visiting professor, College of the Holy Cross, August 25, 1985.

Robert T. Michael: director, National Opinion Research Center, Chicago, September 1, 1984.

Robert N. Mottice, Ashland College: director of taxation and fiscal policy, National Association of Manufacturers, January 14, 1985.

Edgar A. Norton: assistant professor of finance, Northwest Missouri State University, August 13, 1984.

George Sofianos, Harvard University: assistant professor, Graduate School of Business Administration, New York University, September 1984.

Donald L. Westerfield, Southwestern Bell Telephone: adjunct professor of economics, Webster University, September 1, 1984.

#### Leaves for Special Appointments

Elizabeth M. Clayton, University of Missouri-St. Louis: director, Soviet Interview Project, University of Illinois-Champaign, September 1984.

#### NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, use the following style:

A. Please use the following categories (please—do not send public relation releases):

- |   |   |
|---|---|
| 1—Deaths  | 6—New Appointments                                  |
| 2—Retirements                                   | 7—Leaves for Special Appointments (NOT Sabbaticals) |
| 3—Foreign Scholars (visiting the USA or Canada) | 8—Resignations                                      |
| 4—Promotions                                    | 9—Miscellaneous                                     |
| 5—Administrative Appointments                   |   |

B. Please give the name of the individual (SMITH, Jane W.), her present place of employment or enrollment: her new title (if any), new institution and the date at which the change will occur.

C. Type each item on a separate 3×5 card.

D. The closing dates for each issue are as follows: *March*, October 15; *June*, January 15; *September*, April 15; *December*, July 15.

All items and information should be sent to the Assistant Editor, *American Economic Review*, 169 Nassau Street, Princeton, NJ 08542-7067.

---



*Take  
another look!*

New developments in data analysis  
from SPSS and McGraw-Hill

### **NEW** *data analysis facilities*

#### **SPSS<sup>x</sup> User's Guide, second edition**

SPSS Inc. 1986  
1000 pages (tent.)  
(07-046553-3)  
Available in August!

All of the material from the first edition, plus additions through SPSS<sup>x</sup> Release 2.1: probit, logit, and hierarchical loglinear modeling, ALSCAL, new clustering procedures, an all-new plotting procedure, new facilities for reading files from other analysis systems (including SAS and OSIRIS) and DBMS's, and more.

### **SOPHISTICATED** *new display capabilities*

#### **SPSS Graphics**

SPSS Inc. 1985  
320 pages  
(07-046554-1)

A complete guide and reference for the new interactive SPSS Graphics system. This book explains how to use SPSS Graphics on its own or as a powerful extension to SPSS<sup>x</sup>, allowing users to present their data using a wide variety of charts, graphs, maps, and pages of text.

#### **SPSS<sup>x</sup> TABLES**

SPSS Inc. 1985  
224 pages  
(07-046558-4)

Explains how to produce publication-quality tables in a variety of formats using the new optional TABLES procedure in SPSS<sup>x</sup> Release 2.1. Using simple English commands, TABLES allows the user to combine multiple variables in a single table, with flexible options for displaying totals, percentages, and other statistics.

### **MORE** *complete documentation*

#### **SPSS<sup>x</sup> Advanced Statistics Guide**

Marija Norusis 1985  
432 pages  
(07-046548-7)  
Available Now!

A reference for researchers and a text for the multivariate statistics course. This book explains statistical concepts and SPSS<sup>x</sup> operations with interesting examples drawn from actual research.

Topics include: multiple linear regression analysis, discriminant analysis, factor analysis, cluster analysis, multivariate analysis of variance, repeated measures analysis of variance, hierarchical loglinear models, logit models, models for ordinal data, and tests of symmetry.

### **STILL AVAILABLE**

#### **SPSS<sup>x</sup> Basics**

SPSS Inc. 1984  
214 pages  
(07-060524-6)

#### **SPSS<sup>x</sup> Introductory Statistics Guide**

Marija Norusis 1983  
276 pages  
(07-046549-5)

**SPSS Inc.**  
444 N. Michigan Avenue  
Chicago, Illinois 60611



College Division  
McGraw-Hill Book Company  
1221 Avenue of the Americas  
New York, New York 10020

SPSS<sup>x</sup> is a trademark of SPSS Inc.

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

## **Inflation, Stagflation, Relative Prices, and Imperfect Information**

**ALEX CUKIERMAN**

Professor Cukierman presents a summary view of the recent imperfect approach to inflation and its real effects, focusing on two types of informational limitations. The first involves situations in which individuals have asymmetric information about the current general price level and consequently confuse relative and aggregate changes in prices. The second considers models in which individuals cannot distinguish permanent from transitory changes in the economic environment. \$32.50

## **The Economics of Industrial Society**

**MICHIO MORISHIMA**

This study offers new ways of understanding the economic problems of industrialized countries, providing an effective critique of current economic theories and developing an original model of the economics (neoclassical, Marxist, Keynesian) of modern industrial society.

Throughout, the author orients his analysis toward solving problems in the real world and explaining the operations of economic institutions in different countries.

Cloth \$49.50 Paper \$15.95

## **Mathematical Economics**

**Second Edition**

**AKIRA TAKAYAMA**

This is a systematic exposition and survey of mathematical economics which emphasizes the unifying structures of economic theory.

It provides the reader with the technical tools and methodological approaches necessary for undertaking original research, gradually taking him from the elementary level to the frontiers of mathematical economics research.

Cloth about \$34.50 Paper about \$18.95

## **Growth, Acquisition and Investment**

***An Analysis of the Growth of Industrial Firms and  
Their Overseas Activities***

**MANMOHAN S. KUMAR**

Professor Kumar presents the results of an empirical investigation into four aspects of firm growth: the relationship between size, growth and profitability; the degree of trade-off between growth by acquisition and growth by new investment; the role of different forms of financing; and the implications of external markets and overseas production for firms' performance. \$37.50

## **The Philosophy of Economics**

***An Anthology***

**DANIEL M. HAUSMAN, Editor**

This anthology of essays on methodological problems in economics contains the classic discussions by figures such as Mill, Marx, Weber, Veblen, Keynes and many others. It also provides a sampling of work recently undertaken on this topic and includes a comprehensive introduction by the editor.

Cloth \$39.50 Paper \$14.95



**Cambridge  
University  
Press**

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

**A new edition . . .**

**Economic Theory in Retrospect**

**Fourth Edition**

**MARK BLAUG**

This new edition of a classic study in the history of economic thought, focusing on theories rather than theorists, is substantially revised and updated throughout, with major new sections on monetary theory and macroeconomics, and a new chapter on location theory.

From the reviews of previous editions:

"It is a magnificent achievement in the breadth of its scholarship, the incisiveness of its insights, and in the depth of its knowledge of economics past and present."—*The Economic Journal*

Cloth about \$69.50 Paper about \$27.95

**Marshall, Orthodoxy and the Professionalisation of Economics**

**JOHN MALONEY**

John Maloney explores the orthodoxy that economist Alfred Marshall constructed around his own theories and vision of the economics future.

Those who agree with Keynes that "the world is ruled by little else" but the ideas of economists and political philosophers will want to read this book for the light it sheds on how and why, early in this century, one set of economic ideas came to exert a dominance which has persisted to this day.

*Cambridge Studies in the History and Theory of Politics*

About \$47.50

**International Economic Policy Coordination**

**WILLIAM BUITER and RICHARD MARSTON, Editors**

This volume, based on a conference organized jointly by the Centre for Economic Policy Research, examines recent developments in international coordination of economic policy. About \$39.50

**Energy Policy in America Since 1945**

**A Study of Business-Government Relations**

**RICHARD H.K. VIETOR**

"This book should be read by anyone who is interested in energy policy, business history, or business-government relations—which means everyone."—Robert B. Stobaugh, *Harvard Business School*

Since the development of a sound energy policy for cheap fossil fuel—coal, petroleum, and gas—has become a central concern in shaping the economy, we have needed a balanced, authoritative account of the history of energy policy in the US, which, at last, we have in Richard Vietor's book. \$29.95

**An Economic Theorist's Book of Tales**

**Essays that Entertain the Consequences of New Assumptions in Economic Theory**

**GEORGE A. AKERLOF**

This book is a collection of essays exploring the consequences of making non-standard economic assumptions. It breaks with traditional economic theory, which relied upon a tacit and "classical" set of assumptions that have gradually acquired a life of their own in terms of how economists write, think, and justify economic models. Cloth \$29.95 Paper \$8.95

All prices subject to change. Order from your bookstore or call our Customer Service Department at 1-800-431-1580 (outside New York State and Canada). MasterCard or Visa accepted.

**Cambridge University Press • 32 East 57th Street • New York, N.Y. 10022**



# AEA sponsored Group Life Insurance for you and your family— at attractive rates!

The AEA Group Life Insurance Plan can help provide valuable supplementary protection—at attractive rates—for eligible members and their dependents.

Because AEA participates in a large Insurance Trust which includes other scientific and technical organizations, the low cost may be even further reduced by premium credits. In the past eight years, insured members received credits on their April 1 semiannual payment notices averaging over 44% of their annual premium contributions. (These credits are based on the amount paid during the previous policy year ending September 30.) Of course future premium credits, and their amounts, cannot be promised or guaranteed.

Now may be a good time for you to re-evaluate your present coverage and look into AEA Life Insurance. Just fill out and return the coupon for more details at no obligation.

**Administrator, AEA Group Insurance Program**  
1255 23rd Street, N.W.  
Washington, D.C. 20037

G-4

Please send me more information about the AEA Life Insurance Plan.

Name \_\_\_\_\_ Age \_\_\_\_\_

Address \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_ Zip \_\_\_\_\_

Or—call today Toll-Free 800-424-9883  
(Washington, DC area, call 296-8030)

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

## MONEY: how to handle it, understand it . . . make it

*A Modern Classic!*

### THE AFFLUENT SOCIETY

Fourth Revised Edition

By JOHN KENNETH GALBRAITH

MENTOR 0-451-62394-0 \$4.95/\$5.95

### THE MONEY BAZAARS

*Understanding the Banking Revolution Around Us*

By MARTIN MAYER

MENTOR 0-451-62390-8 \$4.95/\$5.95

### THE ZURICH AXIOMS

*Investment Secrets of the Swiss Bankers*

By MAX GUNTHER

PLUME 0-452-25659-3 \$6.95/\$8.75

\*Canadian price Prices subject to change

**NEW AMERICAN LIBRARY**  
1633 Broadway, New York, NY 10019

**NAL**

## OXFORD REVIEW OF ECONOMIC POLICY

### A Major New Economic Journal (1st issue May 1985)

This important new journal is aimed at a wide audience to give informative expert exposition, comment and criticism on the latest economic research.

Each issue contains the latest "Oxford Assessment" appraising recent economic policies and performance, the latest forecast provided by Oxford Economic Forecasting, and a set of articles on a selected theme.

The theme of the first issue is **Monetary and Fiscal Policy** with articles by **Professor David Laidler**, **Professor Marcus Miller**, **Professor Ben Friedman** and **Professor David Hendry**.

#### Introductory Special Offer

Volume One (1985) and Volume Two (1986) can both be obtained at an introductory rate.

**Special Rates:** Institutions U.S. \$140, \*Personal U.S. \$60.

**Subscription Rates Volume One (1985 only):** Institutions: U.S. \$90, \*Personal: U.S. \$45

\*Personal subscriptions can only be sent to a private address

**ORDER FORM: Oxford Review of Economic Policy**

☐ Please record my subscription to Volume One. I enclose my remittance of \_\_\_\_\_

☐ I would like to take advantage of the special introductory offer. Please record my subscription to Volumes One and Two. I enclose my remittance of \_\_\_\_\_

Offer closes end December 1985.

☐ Please send me additional information about **Oxford Review of Economic Policy**.

Name .....

Address .....

Please charge my Access/Visa/Diners/American Express Card. ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

Signature ..... Expiry Date .....

If address registered with credit card company differs from that given above please supply details.

**Oxford Journals**, Oxford University Press, Walton Street, Oxford OX2 6DP, UK

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

IN A NEW EDITION

## **PROGRAMS IN AID OF THE POOR**

FIFTH EDITION

**Sar A. Levitan**

While the Reagan administration's onslaught on the welfare system has not resulted in an appreciable drop in the total level of federal contributions to the poor, the increase in the number of poor—from 31.8 million in 1980 to 35.3 million three years later—has resulted in a sharp decline in per capita aid. In this extensive and timely revision of his classic work, Sar A. Levitan surveys all the existing federally funded programs in aid of the poor and evaluates the impact of the present administration's policies on welfare and assistance programs to America's needy.

**\$17.50 hardcover \$5.95 paperback**

NOW IN PAPERBACK!

## **ALTERNATIVE ROUTES TO FORMAL EDUCATION**

**edited by Hilary Perraton**

The demand for education is outstripping the capacity of many countries to build schools or to recruit and pay teachers. One of the alternatives to the traditional classroom—known as distance teaching—combines correspondence courses with radio or television broadcasts and occasional face-to-face study. This book examines the variety of ways in which distance teaching has been used, provides comparisons of specific cases, analyzes their costs, and considers the effectiveness of distance teaching versus traditional education.

*Published for the World Bank*

**\$14.95 paperback \$35.00 hardcover**

NOW IN PAPERBACK!

## **AGRICULTURAL DEVELOPMENT**

**An International Perspective**

REVISED AND EXPANDED EDITION

**Yujiro Hayami and Vernon W. Ruttan**

A basic reference in the field since its initial publication in 1971, this text here appears in an enlarged and updated new edition—and is made available in paperback for the first time. The authors identify the capacity to develop technology consistent with resource endowments as the single most important variable in the growth of agricultural productivity. In this new edition, they also test their hypothesis against new data and more recent development experience.

*The Johns Hopkins Studies in Development*

**\$18.95 paperback \$39.50 hardcover**



---

**THE JOHNS HOPKINS UNIVERSITY PRESS**

701 West 40th Street, Baltimore, Maryland 21211

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

# BROOKINGS

## **Assessing Tax Reform**

**Henry J. Aaron and Harvey Galper**

The reform of the U.S. tax system is a central political issue. This book offers a concise, nontechnical evaluation to help general readers and students understand the tax reform issues Congress is now debating. The authors identify the major alternative proposals, analyze principles of taxation that can be used for judging them, and present their own program for a fair, efficient, and less complex tax structure.

April 1985/c. 175 pages/\$8.95 paper/\$22.95 cloth

## **Who Paid the Taxes, 1966-85?**

**Joseph A. Pechman**

In this sequel to the widely acclaimed Brookings book *Who Bears the Tax Burden?*, Pechman analyzes how the distribution of federal, state, and local taxes has changed in the past two decades. Presenting estimates based on a unique series of microdata sets, he concludes that the tax system became less progressive between 1966 and 1985, primarily because the corporation income and property taxes declined in importance while heavier emphasis was being placed on the payroll tax.

January 1985/84 pages/\$8.95 paper/\$22.95 cloth

## **Taxes, Loans, and Inflation**

**Eugene Steuerle**

Steuerle examines how the misallocation of capital results from the interaction of tax laws, the operation of the market for loanable funds, and inflation. Focusing on tax arbitrage and inflation-induced discrimination among taxpayers and borrowers, he analyzes the taxation of capital income and discusses several related issues. He concludes with a reform agenda that calls for the adoption of a broader-based, flatter-rate income tax.

May 1985/c. 230 pages/\$9.95 paper/\$26.95 cloth

## **OPEC's Investments and the International Financial System**

**Richard P. Mattione**

After sharply increasing oil prices in late 1973, members of the Organization of Petroleum Exporting Countries built up sizable financial holdings. Mattione is the first to analyze in detail how OPEC nations have used that wealth. This study surveys the size and distribution of each nation's investments; their effects on international financial markets; the financial, political, and developmental motives behind the investment strategies; and the outlook for the late 1980s.

January 1985/202 pages/\$9.95 paper/\$26.95 cloth

## **The Global Factory: Foreign Assembly in International Trade**

**Joseph Grunwald and Kenneth Flamm**

Firms in industrial countries are increasingly shifting labor-intensive production processes abroad to developing countries with an abundance of cheap labor. In this evaluation of foreign assembly, the authors first examine the semiconductor industry and then present case studies of the assembly industry in Mexico, Haiti, and Colombia. They analyze the domestic, political, social, and economic effects of the reorganization of industry abroad and discuss the policy implications for the United States and its manufacturing partners.

March 1985/259 pages/\$10.95 paper/\$29.95 cloth

### **The Brookings Institution**

1775 Massachusetts Avenue, N.W.  
Washington, D.C. 20036  
(202) 797-6258

# BROOKINGS

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

# **The Journal of International Economics and Economic Integration Offers**

## **\$5,000**

### **For the First Annual Daeyang Prize in Economics**

- The Journal of International Economics and Economic Integration is published biannually by the Institute for International Economics, King Sejong University, Seoul, Korea.
- The purpose of the Journal of International Economics and Economic Integration is to support and encourage research in the area of international trade, international finance and other related economic issues that include general professional interest in international economic affairs.
- The Journal of International Economics and Economic Integration welcomes unsolicited manuscripts, which will be considered for publication by the Editorial Board.
- The Editorial Board will choose fourteen manuscripts for publication on an annual basis.
- The Editorial Board will choose the best manuscript out of the fourteen to be awarded the \$5,000 Daeyang Prize in Economics.
- The manuscripts, which should be accompanied by an abstract of no more than 100 words, should be typewritten, double-spaced, in English with footnotes, references, figures, tables and any other illustrative material on separate sheets.
- Three copies of the manuscript and all accompanying material should be submitted to the following address by October 31, 1985 for consideration for the 1986 publication.

**Institute for International Economics  
King Sejong University  
Seongdong-Ku, Seoul, Korea**

*New from MIT*

## **The Management Challenge**

Japanese Views

*edited by Lester C. Thurow*

The original contributions in this book present Japanese management as the Japanese see it. They show how an economy that has outperformed ours during the last 30 years works. Thurow's introductory essay and his comments on each contribution provide a unifying framework while pointing up the implications that each chapter raises for the reblending of the American economic mixture.

\$14.95

## **Market Structure and Foreign Trade**

*Elhanan Helpman and Paul R. Krugman*

Relating current theoretical work to the main body of trade theory, this book offers entirely new material on contestable markets, oligopolies, welfare, and multinational corporations, and new insights on external economies, intermediate inputs, and trade composition.

\$22.50

## **Deregulating the Airlines**

*Elizabeth E. Bailey, David R. Graham, and Daniel P. Kaplan*

"An excellent evaluative chronicle of perhaps the most significant regulatory reform of our time."—James C. Miller III, Chairman, Federal Trade Commission

\$19.95

## **Antitrust and Regulation**

Essays in Memory of John J. McGowan

*edited and introduced by Franklin M. Fisher*

This collection of original essays by economists and lawyers addresses such issues as the U.S. government's merger guidelines, antitrust in regulated industries, the connection between profitability and market share, and the question of what constitutes anticompetitive behavior.

\$35.00

*Now available in paperback*

## **Folded, Spindled, and Mutilated**

Economic Analysis and U.S. vs. IBM

*Franklin M. Fisher, John J. McGowan,  
and Joen E. Greenwood*

"A rigorous analysis of competition in the computer industry."

—Edwin McDowell, *The New York Times*

\$9.95

## **The MIT Press**

28 Carleton Street Cambridge, MA 02142

# JOURNAL OF ACCOUNTING & ECONOMICS

Published by North-Holland in collaboration with The Graduate School of Management, The University of Rochester

## Editors:

ROSS L. WATTS and JEROLD L. ZIMMERMAN,  
Graduate School of Management, University of  
Rochester, Rochester, NY 14627, USA

## Consulting Editor:

ROBERT S. KAPLAN,  
Graduate School of Industrial Administration,  
Carnegie-Mellon University, Pittsburgh, PA, USA.

The *Journal of Accounting and Economics* encourages the application of economic theory to the explanation of accounting phenomena. The theories of the firm, public choice, government regulation, and agency theory, in addition to financial economics, can contribute significantly to increasing our understanding of accounting. The *JAE* provides a forum for the publication of the highest quality manuscripts which employ economic analyses of accounting problems. A wide range of methodologies are encouraged and covered:

- the determination of accounting standards
- government regulation of corporate disclosure
- the information content and role of accounting numbers in capital markets
- the role of accounting in financial contracts and in monitoring agency relationships
- the theory of the accounting firm
- government regulation of the accounting profession
- statistical sampling and the loss function in auditing
- the role of accounting within the firm (managerial accounting)

## CONTENTS OF SPECIAL VOLUME 7 (March 1985):

### Management Compensation and the Managerial Labor Market

Kevin J. Murphy, Corporate Performance and Managerial Remuneration, An Empirical Analysis; Anne T. Coughlan and Ronald M. Schmidt, Executive Compensation, Management Turnover, and Firm Performance: An Empirical Investigation; George J. Benston, The Self-Serving Management Hypothesis: Some Evidence; Paul M. Healy, The Effect of Bonus Schemes on Accounting Decisions; Robert S. Kaplan, Comments on Paul Healy, "The Effect of Bonus Schemes on Accounting Decisions"; James A. Brickley, Sanjai Bhagat and Ronald C. Lease, The Impact of Long-Range Managerial Compensation Plans on Shareholder Wealth; Hassan Tehrani and James F. Waageleijn, Market Reaction to Short Term Executive Compensation Plan Adoption; Jerold B. Warner, Stock Market Reaction to Management Incentive Plan Adoption: Overview; W. Bruce Johnson, Robert P. Magee, Nandu J. Nagarajan and Harry A. Newman, An Analysis of the Stock Price Reaction to Sudden Executive Deaths: Implications for the Managerial Labor Market; G. William Schwert, A Discussion of C.E.O. Deaths and the Reaction of Stock Prices; Richard A. Lambert and David F. Larcker, Golden Parachutes, Executive Decision-Making, and Shareholder Wealth; Sherwin Rosen, Commentary on "Golden Parachutes, Executive Decision-Making, and Shareholder Wealth"; Wilbur Lewellen, Claudio Loderer and Ahron Rosenfeld, Merger Decisions and Executive Stock Ownership in Acquiring Firms; Wayne H. Mikkelsen and Richard S. Ruback, Takeovers and Managerial Compensation: A Discussion; Artur Raviv, Management Compensation and the Managerial Labor Market: An Overview; William H. Meckling, Three Reflections on Performance Rewards and Higher Education.

## Subscription Information:

Volume 7 is scheduled for publication in March 1985.  
Subscription price: Dfl. 185.00/US \$68.50.  
Add Dfl. 22.00/US \$8.25 for handling and postage.  
**Total price: Dfl. 207.00/US \$76.75.**  
ISSN 0-165-4101

## ORDER FORM

Send this form or a photocopy to:  
**ELSEVIER SCIENCE PUBLISHERS**  
Subscription Order Dept.,  
P.O. Box 211, 1000 AE Amsterdam,  
The Netherlands

In the U.S.A. and Canada:  
**ELSEVIER SCIENCE PUBLISHERS**  
Attn: Journal Information Center,  
52 Vanderbilt Avenue, New York,  
NY 10017, U.S.A.

Orders from individuals should be  
accompanied by a remittance.

**NORTH-HOLLAND**

Please enter my subscription to the *Journal of Accounting & Economics* (1985: Volume 7) at Dfl. 207.00/US \$76.75 incl. handling & postage.

☐ Payment enclosed (Bank Draft/Eurocheque/International Money Order/Personal Cheque/Postal Cheque/Official Purchase Order Form)

☐ Charge my credit card: ☐ Access ☐ Eurocard ☐ Mastercard  
☐ American Express ☐ VISA

Card No. \_\_\_\_\_ Valid until \_\_\_\_\_

Name \_\_\_\_\_

Address \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

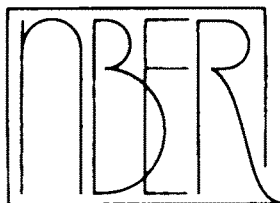
Date \_\_\_\_\_ Signature \_\_\_\_\_

The Dutch guildler prices are definitive.

0907NH/JRNL/ECON

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

# CHICAGO



Published for the  
**National Bureau of  
Economic Research**

## **CORPORATE CAPITAL STRUCTURES IN THE UNITED STATES**

*Edited by Benjamin M. Friedman*

Continuing the project begun in *The Changing Roles of Debt and Equity in Financing U.S. Capital Formation*, this volume focuses on the financial side of capital formation, particularly the issue of financing capital through debt and equity. It is a useful compendium on corporate finance for those involved in decisions on financing corporate investment as well as academics interested in the financial structure of U.S. corporations and the relationship of corporate needs for capital to the needs of the government.

*An NBER Project Report*  
Cloth \$48.00 408 pages

## **FEDERAL TAX POLICY AND CHARITABLE GIVING**

*Charles T. Clotfelter*

In this study, Clotfelter demonstrates that changes in tax policy — effected through legislation or inflation — can have a significant influence on the level and composition of charitable giving. Clotfelter focuses on empirical analysis of the effects of tax policy on charitable giving in four major areas: individual contributions, volunteering, corporate giving, and charitable bequests. The result is a model for policy-oriented research efforts as well as a timely contribution to the evidence that must inform proposals for tax reform.

*An NBER Monograph*  
Cloth \$39.00 336 pages

## **THE INTERNATIONAL TRANSMISSION OF INFLATION**

*Michael R. Darby, James R. Lothian,  
Arthur E. Gandolfi, Anna J. Schwartz,  
and Alan C. Stockman*

How does inflation catch fire and spread from country to country? This book offers comprehensive answers, including the controversial argument that the United States, through its policy of monetary growth, was the primary instigator of inflation both at home and abroad during the 1970s. The authors draw their conclusions from a multicountry data base and a model of international transmission more complete than any other yet constructed.

*An NBER Monograph*  
Paper \$22.50 744 pages

## **SOCIAL EXPERIMENTATION**

*Edited by Jerry A. Hausman and  
David A. Wise*

Since 1970 the United States government has spent over half a billion dollars on social experiments intended to assess the effect of potential tax policies, health insurance plans, housing subsidies, and other programs. The contributors to this volume consider, among other questions, whether these experiments were worth their cost, whether they revealed information that could not have been learned by other means, and whether the experiments could have been better designed and analyzed.

*An NBER Conference Report*  
Cloth \$33.00 308 pages

The University of **CHICAGO** Press

5801 South Ellis Avenue, Chicago, IL 60637

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers



# Chart new economic trends for your students



## **International Trade and Lending**

**MICHAEL CONNOLLY**

A succinct yet complete examination of international trade and lending theory. The book also incorporates into the body of trade theory problems involving borrowing and lending over time.

**144 pp. May 1985 \$28.95 ISBN 0-03-071166-5**

## **Barter in the World Economy**

edited by **BART S. FISHER** and  
**KATHLEEN M. HARTE**

This new book, consisting of 11 incisive commentaries on the subject of barter, addresses its role and counter-trade in today's economy, examines its mechanics, and explains why barter has become an attractive alternative to traditional currency transactions.

**288 pp. (tent.) May 1985 \$37.95 (tent.)**  
**ISBN 0-03-071609-8**

# **An Introduction to Airline Economics**

**THIRD EDITION**

**WILLIAM W. O'CONNOR**

This latest revision of this popular introduction covers the many changes that have occurred in the airline industry since the rise of the deregulation movement in the mid-seventies.

**224 pp.**

**April 1985**

**\$24.95**

**ISBN 0-03-001827-7**

## **Also of interest . . .**

### **Investment Incentives and Performance Requirements**

PATTERNS OF INTERNATIONAL TRADE, PRODUCTION, AND INVESTMENT

**STEPHEN E. GUISINGER AND ASSOCIATES**

**336 pp. April 1985 \$35.95**  
**ISBN 0-03-002443-9**

### **Teachers Unionism and Its Impact**

A STUDY OF CHANGE OVER TIME  
**DOROTHY KENT JESSUP**

**208 pp. (tent.) June 1985**  
**\$29.95 (tent.)**

**ISBN 0-03-002858-2**

### **The Economics of R & D Policy**

edited by **GEORGE S. TOLLEY, JAMES H. HODGE, and JAMES F. OCHMKE**

**192 pp. (tent.) June 1985**  
**\$28.95 (tent.)**

**ISBN 0-03-000892-1**

### **Regional Growth and Decline in the United States**

SECOND EDITION

**BERNARD L. WEINSTEIN, HAROLD T. GROSS, and JOHN REES**

**160 pp. (tent.) June 1985**  
**\$25.95 (tent.)**

**ISBN 0-03-062044-9**

### **The Economics of the Caribbean Basin**

edited by **MICHAEL CONNOLLY and JOHN McDERMOTT**

**304 pp. (tent.) May 1985**  
**\$36.95 (tent.)**

**ISBN 0-03-001674-6**

**Order from your local bookseller or direct from**

# **PRAEGER**

**PUBLISHERS**

**521 Fifth Avenue, New York, New York 10175**

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*



Begin making plans to attend the

# **Annual Meeting of The American Economic Association, Centennial Celebration**

(in Conjunction with Allied Social Science Associations)  
to be held in

## **NEW YORK, NY**

**Dec. 28-30, 1985**

---

The Employment Center opens Friday, December 27.

---

See the Notes section of the September *AER* for the American Economic Association's preliminary program.

---

The 1986 meeting will be held in New Orleans, LA, December 28-30.

---

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*



# SCIENTIFIC WORD PROCESSING AS EASY AS $\pi$

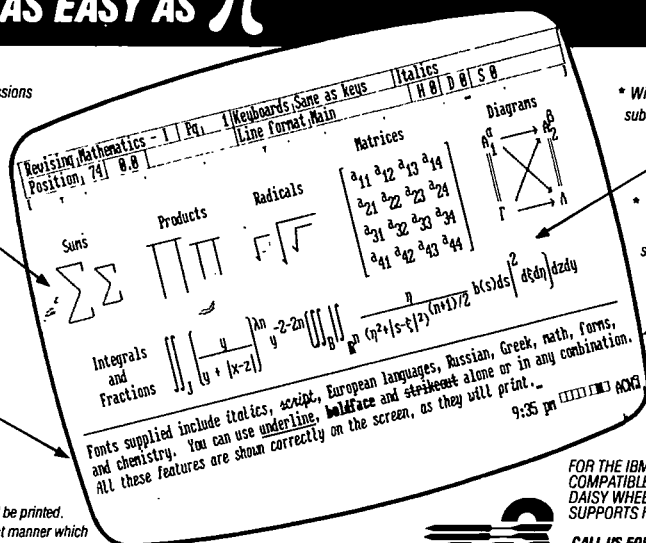
\* With T<sup>3</sup>, save complex expressions by name, and you never have to type them again.

\* With T<sup>3</sup>, define and use up to 1024 characters in a single document.

With T<sup>3</sup> complex expressions appear on the screen as they will be printed. You enter them in a simple, direct manner which won't interfere with your train of thought. You can compose scientific manuscripts directly at the keyboard.

\* With T<sup>3</sup>, use up to 25 levels of subscripting and superscripting.

\* With T<sup>3</sup>, format text directly on the screen, with line spacing, underline, boldface, and italics all visible.



Fonts supplied include italics, script, European languages, Russian, Greek, math, forms, and chemistry. You can use underline, boldface and strikethrough alone or in any combination. All these features are shown correctly on the screen, as they will print.

**THE SCIENTIFIC WORD PROCESSING  
SYSTEM THAT'S EASY TO USE!**



T<sup>3</sup> SOFTWARE  
RESEARCH, INC.

1190-B FOSTER ROAD - LAS CRUCES, NEW MEXICO 88001

FOR THE IBM PC, XT, AT AND MANY  
COMPATIBLES. SUPPORTS: DOT MATRIX,  
DAISY WHEEL AND LASER PRINTERS.  
SUPPORTS HERCULES GRAPHICS CARD

CALL US FOR MORE INFORMATION  
1-800-874-2383  
IN NEW MEXICO (505)522-4600  
TELEX: 317629

## N E B R A S K A

### Production, Income, and Welfare

The Search for an Optimal Social Order

By Jan Tinbergen

*Production, Income, and Welfare* focuses on the recurring theme of Tinbergen's work, drawing together his various concerns: using basic tools of economic and econometric analysis, Tinbergen attempts to discover scientifically the best socioeconomic policy and the optimum mix of the market and planned elements of the economy.

Jan Tinbergen was the first economist to receive the Nobel Prize in Economics (1969), a just reward for his contributions to the discipline. His work has ranged widely, covering such diverse areas as econometrics, business cycles, income distribution, and production functions, and in each he has been responsible for important advances.



UNIVERSITY OF NEBRASKA PRESS

901 North 17th Lincoln, NE 68588

ca. 256 pages.

ISBN 0-8032-4412-6 \$21.95

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

# SPREADSHEET STATISTICAL PACKAGE

## ySTAT™

Mainframe Capability, Accuracy and Speed  
for Your IBM Personal Computer

### WORKS LIKE LOTUS' 1-2-3, WITH FEATURES LIKE SAS

- ySTAT combines the best of microcomputer (Lotus 1-2-3) and the best of mainframe (SAS) in ONE. Yes, in ONE.
- It is almost effortless to use ySTAT even for the first time. And if you are familiar with Lotus 1-2-3 or any other spreadsheet program, you are ready to start using ySTAT. As in 1-2-3, you simply move the cursor to the appropriate menu or submenus and make your selection.
- Lets you enter or read data into the spreadsheet so that you can see the data on the screen by scrolling sideways or up and down the spreadsheet. You may enter, edit, copy, move, delete, select or do whatever you like to the data.
- ySTAT features advanced windowing technique for selecting variable names and displaying outputs and tables.
- Uses dynamic memory allocation to give you the maximum amount of memory for each procedure (PROC) that you may use at any time. You do not have to swap disks all the time.
- Accepts data received from time sharing systems, or data generated by other PC programs. It reads individual raw data for up to 255 variables per record as well as multi-way tables as input. The data may be numeric, alphabetic, or hybrid, with or without missing values.
- ySTAT simplifies data transformation by performing vector matrix operations from cell formula that only one formula is required for creating a whole column of values for a new variable.
- It includes the following procedures (PROC): Means, Frequency, Summary, Correlate, Crosstab, Table, ANOVA, OLS, WLS, 2SLS, Autoregression, Pooling of time series and Cross section, etc.
- ySTAT is a fast, highly accurate, economical, and, in many ways, superior alternative to main frame computing. Unlike the mainframe and many of its micro-offsprings which display a large data set or table in fragmented parts and line by line, ySTAT's screen displays are rapidly presented as screen pages and can be scrolled or flipped over in all directions with a touch of one or two keys. A large table or data can be viewed continuously.

#### SYSTEM REQUIREMENTS

IBM PERSONAL COMPUTER (PC, XT, AT), or other MS DOS compatibles; 1 floppy disk drive or a hard disk system; DOS 2.0 or 3.0 and 256K memory; with or without 8087/80287 math chip.

PRICE: \$395. Demo disk available at \$5.

Introductory price at \$295 for the first 1000 inquirers (including demo purchasers).

### YOU'LL BE GLAD THAT YOU INQUIRED

#### MING TELECOMPUTING INC.

Telecommunications and Statistics for Microcomputers

23 Oak Meadow Road, P.O. BOX 101, Lincoln Center, MA 01773 (617) 259-0391

MING TELECOMPUTING INC., P.O. BOX 101, Lincoln Center, MA 01773 (617) 259-0391

Please send: ( ) ySTAT Package for \$295. (Massachusetts residents: please add \$14.95)  
( ) ySTAT demo disk and information/sample output for \$5 (to be credited for purchase)  
( ) ySTAT information/sample output.  
( ) other \_\_\_\_\_  
My system: ( ) IBM PC. ( ) IBM XT. ( ) IBM AT. Memory \_\_\_\_\_ K  
( ) other \_\_\_\_\_  
( ) with 8087 or 80287 math co-processor.  
Payment: ( ) a check is enclosed.  
( ) Visa. ( ) MasterCard. Credit Card # \_\_\_\_\_ Expiration Date: \_\_\_\_/\_\_\_\_  
( ) a university or government purchase order enclosed.

Name \_\_\_\_\_ Telephone: ( ) \_\_\_\_\_ Signature (if charged) \_\_\_\_\_

Address \_\_\_\_\_

(SAMPLE OUTPUTS ARE AVAILABLE UPON REQUEST.)

IBM is a registered trademark of International Business Machines Corp., 1-2-3 of Lotus, SAS of SAS Institute

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

**ROBERT A. DAHL**

## **A Preface to Economic Democracy**

Dahl, author of the classic, *A Preface to Democratic Theory*, claims we can achieve a society of true social and political equality without sacrificing liberty by making a radical extension of democracy to the economic order. He presents an empirically informed and philosophically acute defense of "work-place democracy" that challenges our common assumptions and makes the startling proposal that only a radical transformation of American social vision can truly reconcile the demands of liberty, justice, and efficiency. \$14.95

**JOHN M. QUIGLEY & DANIEL L. RUBINFELD, Editors**

## **American Domestic Priorities**

### **An Economic Appraisal**

High deficits, large military and social security expenditures, and the "New Federalism" have put the future of many domestic programs in doubt. Twenty-six nationally prominent economists address questions fundamental to the design of education, housing, transportation, environmental, and anti-poverty policy, proposing imaginative alternatives to further budget cuts for reducing the federal deficit. \$32.50 cloth, \$9.95 paperback

**LINDA ALEXANDER RODRIGUEZ**

## **The Search for Public Policy**

### **Regional Politics and Government Finances in Ecuador, 1830-1940**

Based on meticulous research into Ecuadorian, U.S., and British archives, this study of the evolution of an active state in Ecuador during the first century of its national existence cuts through the political and ideological propaganda that has obscured the country's history. \$32.50

**LORNA M. DANIELLS**

## **Business Information Sources**

### **New, Revised Edition**

Here is the long-awaited revised edition of this essential guide to business books and reference sources. Aimed at the practicing business person as well as the business student and the librarian, this guide has been hailed as "an absolutely indispensable book for any library that ever has a business question" (*Choice*). \$24.95

**M. I. FINLEY**

## **The Ancient Economy**

### **Second Edition, new in paperback**

This book examines the various aspects of what we call the economy of the Graeco-Roman world with the concepts appropriate to their behavior and values. \$8.95

---

At bookstores or order from

**University of California Press**

**BERKELEY 94720**

# Centennial T-Shirt



celebrating the  
100th Anniversary  
of the

**American  
Economic Association**  
1885 - 1985.

Now available by mail.

Your next chance: the year 2085

- Collector's Item
- Limited Edition
- Satisfaction Guaranteed

Top quality name brand T-Shirt (50% cotton and 50% polyester for easy care and no shrinkage) imprinted with the familiar AER cover design in authentic burgundy red with the AEA Seal in black, printed over your choice of light blue or tan fabric.

PLEASE PRINT CLEARLY

Send to:  
**ACADEMICS**  
1800 E. Capitol Drive  
Suite 2F  
P.O. Box 11768  
Milwaukee, WI 53211

Send \_\_\_\_\_ AEA's Centennial T-Shirts at \$9.95 each including handling and delivery.

Enclosed: check \_\_\_\_\_ money order \_\_\_\_\_ payable to ACADEMICS.

Or charge my MasterCard \_\_\_\_\_ VISA \_\_\_\_\_.

Card # \_\_\_\_\_ Exp. Date \_\_\_\_\_

Size Chart	
Small	34-36
Medium	38-40
Large	42-44
X-Large	46-48

Name \_\_\_\_\_

Address \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_ Zip \_\_\_\_\_

Signature \_\_\_\_\_

Quantity	Size	Color

Wisconsin residents add 5% sales tax. Allow 2-3 weeks for delivery.

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

# ENERGY FACTS AND FIGURES

Robert H. Romer, Amherst College

A concise and useful energy handbook containing the essential data that anyone needs when dealing with the energy problem—72 pages of tables and graphs showing the history of energy production and consumption in the world and the United States from 1850 to 1983, with a detailed examination of energy use in the United States in 1983.

Carefully compiled by Robert H. Romer, Professor of Physics at Amherst College, and based in part on the extensive appendices to his earlier energy book, **Energy Facts and Figures** includes world-wide and U.S. electrical generating capacity, per capita GNP and power consumption figures since 1900, consumer prices of common sources of energy, solar energy and degree-day data, the energy content of various fuels, the dimensions and composition of the earth, its atmosphere and its oceans, the earth's water cycles and energy flows, energy requirements for various modes of transportation and for common electrical appliances, an amortization table for comparing capital and operating costs, and more. An extensive table of conversion factors is included.

\$8.95 postpaid. ISBN 0-931691-17-6 (paper)

**Spring Street Press**

104 Spring St., Amherst, Mass. 01002

---

---

## INTERNATIONAL MONETARY FUND STAFF PAPERS

The March issue of the quarterly theoretical journal of the International Monetary Fund contains the following articles:

*Balance of Payments Financing and Reserve Creation*

Rudolf R. Rhomberg

*Economic Stabilization in Planned Economies: Toward an Analytical Framework*

Thomas A. Wolf

*Trade and Financial Liberalization Given External Shocks and Inconsistent Domestic Policies*

Mohsin S. Khan and Roberto Zahler

*Export Demand and Supply for Groups of Non-Oil Developing Countries*

Marian E. Bond

*Exchange Rate Dynamics and Intervention Rules*

A. Blundell-Wignall and P.R. Masson

*The Changing Role of International Bank Lending in Development Finance*

David Folkerts-Landau

Subscriptions for *Staff Papers* (US\$15.00 for a volume, US\$4.00 for a single issue, special rates—US\$7.50 a volume—to university libraries, faculty, and students) should be sent to:

Publications Unit, Box E-199  
International Monetary Fund  
700 19th Street, N.W.  
Washington, D.C. 20431, U.S.A.  
Telephone (202) 473-7430

---

---

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers



The American Economics Association (AEA) is now soliciting applications to host the AEA Summer Program for Minority Students, for three summers beginning in 1986.

This program is now in its twelfth year and is currently at the University of Wisconsin at Madison. Previous host institutions have been Berkeley, Northwestern, and Yale.

The intent of the program is to increase the number of Blacks, Hispanics, and Native Americans pursuing the Ph.D. in Economics. In recent years the course of study has been an intensive eight to ten week program in intermediate microeconomics and macroeconomics, at the honor's level, and courses in econometrics or mathematical economics.

Funding for the program has been provided by the hosting institutions and grants to the AEA. Applications should be sent to Professor Donald J. Brown, chairman of the AEA Committee on the Status of Minority Groups in the Economics Profession, CSMGEP, no later than September 1, 1985.

CSMGEP  
Attention: Donald J. Brown  
Chairman, Department of Economics  
Yale University  
Box 1972 Yale Station  
New Haven, CT 06520-1972

## ANNOUNCING A NEW STATISTICAL PACKAGE FOR YOUR PERSONAL COMPUTER!

- Ordinary Least Squares Estimation
- Polynomially Distributed Lag Estimation
- Two Stage Least Squares Estimation
- Cochrane-Orcutt estimation feature in each of above
- Usual summary statistics and more including D.W. autocorrelation coeff., log of likelihood, correlation matrix, var-cov matrix for estimated coefficients, etc.
- Prediction with confidence and prediction intervals in OLS
- Handles large numbers of variables and observations

Price: \$99.50 US/\$129.50 CAN\*

Please send me a CAN-AM STATISTICAL PACKAGE. ( ) The payment is enclosed.  
( ) Please charge \$129.50 CAN to my ( ) VISA or ( ) MasterCard account.

Credit Card # \_\_\_\_\_: Expiry date \_\_\_\_\_

Name \_\_\_\_\_ Signature (if charged) \_\_\_\_\_

Address \_\_\_\_\_

\*Ontario residents: Please add 7% provincial sales tax.

Circle one: IBM (PC, PC XT, PC jr); Apple (II+, IIe IIc); Commodore 64

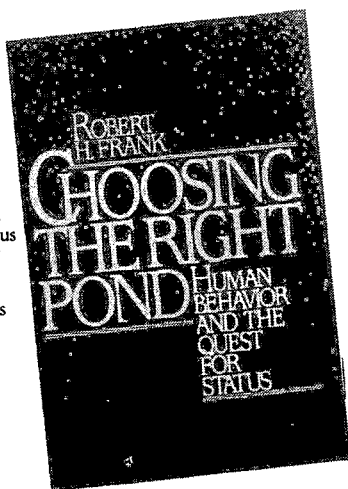
To order please mail this coupon (or photocopy). Make cheque or money order payable to:



**CAN-AM FINANCIAL CONSULTING**  
177 Caddy Ave., Sault St. Marie, Ont. P6A 6H7 CANADA

## Is it better to be a big frog in a small pond or a small frog in a big pond?

Robert H. Frank argues that people's concerns about status affect a surprisingly broad range of economic choices, from salaries to laws and regulations.



"*Choosing the Right Pond* is bound to change the way economists, philosophers, and others think about a great variety of topics ranging from safety regulations and consumers' savings to general principles of distributive justice and social contract theory . . . the book is a landmark in social thought."

David Braybroke  
Dalhousi University, Canada

"Frank is obviously onto something fundamental in his treatment of envy and status as factors in human behavior . . . I couldn't put *Choosing the Right Pond* down."

E.O. Wilson  
author, *On Human Nature*

"It is a rare combination: a book that is profound and fun to read."

James M. Buchanan  
author, *Calculus of Consent*

\$22.95  
at better bookstores,  
or send your check to

**Oxford University Press**

Department HS 200 Madison Ave. New York, NY 10016

Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

## Computer Access to Articles in the JEL Subject Index

On-line computer access to the *JEL* and *Index of Economic Articles* data base of journal articles is currently available through DIALOG Information Retrieval Service. DIALOG file 139 (*Economic Literature Index*) contains complete bibliographic citations to articles from some 260 journals listed in the quarterly *JEL* issues from 1969 through the current issue. The abstracts in *JEL* since June 1984 are also available as part of the full bibliographic record. The file may be searched using free-text searching techniques or author, journal, title, geographic area, date, and other descriptors, including descriptor codes based on the *Index's* four-digit subject classification numbers.

DIALOG offers a variety of contract choices, including the option to pay for only what you use—*no minimum, no initiation or start-up fee*. Most university libraries already subscribe to DIALOG. For information on the DIALOG service, contact your librarian or write to or call:

DIALOG Information Services, Inc., Marketing Department, 3460 Hillview Avenue, Palo Alto, California 94304 (800/227-1927 or 800/982-5838, in California, or 415/858-3785).



**Yes ! Please send me for my library:**

**NATIONAL ECONOMIC PLANNING: *What Is Left?*** \$ \_\_\_\_\_

Don Lavoie / *A Cato Institute Book* • August — ca 336 pages — \$25.00

**BANKING DEREGULATION AND THE NEW COMPETITION** \$ \_\_\_\_\_

**IN FINANCIAL SERVICES**

Kerry Cooper and Donald R. Fraser • Available — 210 pages — \$32.00

**THE FUTURE OF SMALL BANKS IN A DEREGULATED** \$ \_\_\_\_\_

**ENVIRONMENT**

Donald R. Fraser and James Kolari • June — ca 200 pages — \$25.00

**MONITORING GROWTH CYCLES IN MARKET-ORIENTED** \$ \_\_\_\_\_

**COUNTRIES**

*Developing and Using International Economic Indicators*

Philip A. Klein and Geoffrey H. Moore / *A National Bureau of Economic*

*Research Book* • July — ca 352 pages — \$35.00

**EXCHANGE RATES, TRADE, AND THE U.S. ECONOMY** \$ \_\_\_\_\_

Sven W. Arndt, Richard J. Sweeney, and Thomas B. Willett, editors

*An American Enterprise Institute Book* • May — 328 pages — \$39.95

**REVITALIZING AMERICAN INDUSTRY: *Lessons from Our Competitors*** \$ \_\_\_\_\_

Milton Hochmuth and William Davidson, editors •

Available — 420 pages — \$39.95 **Subtotal** \$ \_\_\_\_\_

☐ payment enclosed ☐ bill me

charge my ☐ MC ☐ VISA ☐ AMX

**Massachusetts 5% Sales Tax** \$ \_\_\_\_\_

**Postage and handling (\$1.50/bk)** \$ \_\_\_\_\_

Card No. \_\_\_\_\_ Exp. date \_\_\_\_\_

**TOTAL** \$ \_\_\_\_\_

Signature \_\_\_\_\_

**Send to:**

NAME \_\_\_\_\_

ZIP \_\_\_\_\_

**When you place your order by phone  
please tell the operator you code is AAER685.**

**BALLINGER**  
PUBLISHING COMPANY  
Order Department  
2350 Virginia Avenue, Hagerstown, MD 21740  
To order, call toll free 1-800-638-3030.



Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers

**AMERICAN ECONOMIC ASSOCIATION**  
**1985 ANNUAL MEMBERSHIP RATES**

**Membership includes:**

—a subscription to both *The American Economic Review* (quarterly) plus *Papers and Proceedings* and the *Journal of Economic Literature* (quarterly).

- Regular members with annual incomes of \$30,000 or less ..... \$35.00
- Regular members with annual incomes above \$30,000 but no more than \$40,000 ..... \$42.00
- Regular members with annual incomes above \$40,000 ..... \$49.00
- Junior members (available to registered students for three years only).

Student status must be certified by your major professor or school registrar ..... \$17.50

- In Countries other than the U.S.A., Add \$11.00 to cover postage.
- Family members (persons living at the same address as a regular member, additional memberships without subscription to the publications of the Association) ..... \$7.00

Please begin my issues with:

☐ **March**

☐ **June**  
(Includes *Papers and Proceedings*)

☐ **September**

☐ **December**

First Name and Initial	Last Name	Suffix
Address Line 1		
Address Line 2		
City		
State or Country	Zip/Postal Code	

**MAJOR FIELDS (TWO ONLY)**

LIST FIELDS WITH WHICH YOU CURRENTLY IDENTIFY. SELECT FIELD CODE FROM JEL, "Classification System for Books."

PLEASE TYPE OR PRINT INFORMATION ABOVE; PLEASE SEND CHECK OR MONEY ORDER PAYABLE IN U.S. DOLLARS. CANADIAN AND FOREIGN PAYMENTS MUST BE IN THE FORM OF A U.S. DOLLAR DRAFT ON A NEW YORK BANK.

Endorsed by (AEA member) \_\_\_\_\_

**Below for Junior Members Only**

I certify that the person named above is enrolled as a student at \_\_\_\_\_

\_\_\_\_\_  
Authorized Signature

PLEASE SEND WITH PAYMENT TO:

**AMERICAN ECONOMIC ASSOCIATION**  
1313 21ST AVENUE SOUTH, SUITE 809  
NASHVILLE, TENNESSEE 37212-2786  
U.S.A.



Economics Institute  
1030 13th Street  
Boulder, Colorado 80302 U.S.A.

**GUIDE TO GRADUATE STUDY IN ECONOMICS,  
AGRICULTURAL ECONOMICS, AND DOCTORAL  
DEGREES IN BUSINESS AND ADMINISTRATION**

in the United States of America and Canada, 7th edition, 544 pages

edited by Wyn F. Owen and Larry R. Cross

Published by the Economics Institute—a nonprofit educational corporation sponsored by the American Economic Association and endorsed by the American Agricultural Economics Association.

The **GUIDE** is an indispensable reference book for prospective graduate students—both domestic and foreign—and their advisors and sponsors. Comparative analyses of the programs are given. Over three hundred individual programs are described.

-----  
**ORDER FORM**

Economics Institute  
Publications Center  
1030 13th Street  
Boulder, Colorado 80302 U.S.A.

Please send me \_\_\_\_\_ copy(ies) of the **GUIDE** at \$33.00 per copy. For foreign orders, please enclose an additional \$3.00 for shipping and handling.

\_\_\_\_\_ I enclose \$\_\_\_\_\_ (check or international money order)

\_\_\_\_\_ Please bill me \$\_\_\_\_\_.

\_\_\_\_\_ Charge \$\_\_\_\_\_ to my \_\_\_\_\_ Mastercard, \_\_\_\_\_ Visa, or

\_\_\_\_\_ American Express      Number \_\_\_\_\_

Expires \_\_\_\_\_ Authorized Signature \_\_\_\_\_

For faster service on credit card orders only, call 303-492-8417 ext. 23.

Name \_\_\_\_\_ Title \_\_\_\_\_

Address \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_

Zip Code \_\_\_\_\_ Country \_\_\_\_\_  
(plus four)

# JOB OPENINGS FOR ECONOMISTS

Available only to AEA members and institutions that agree to list their openings.

## Annual Subscription Rates

U.S.A., Canada, and Mexico (first class): \$15.00, regular AEA members and institutions  
\$ 7.50, junior members of AEA  
All other countries (air mail): \$22.50, regular AEA members and institutions  
\$15.00, junior members of AEA

Please begin my issues with:

☐ February ☐ April ☐ June ☐ August ☐ October ☐ December

Name \_\_\_\_\_

First

Middle

Last

Address \_\_\_\_\_

City

State/Country

Zip/Postal Code

Check one:

- ☐ I am a member of the American Economic Association.  
☐ I would like to become a member. My application and payment are enclosed.  
☐ (For institutions) We agree to list our vacancies in JOE.

Send payment (U.S. currency only) to:

THE AMERICAN ECONOMIC ASSOCIATION  
1313 21st Avenue South  
Nashville, Tennessee 37212

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*

# The Economics Institute

Boulder, Colorado

---

- PREPARATION FOR MASTERS AND DOCTORAL DEGREE PROGRAMS IN ECONOMICS, BUSINESS, AND ADMINISTRATION.
  - POSTGRADUATE DIPLOMA PROGRAMS IN RELATED SPECIALIZATIONS.
  - AN ESTABLISHED REPUTATION FOR ACADEMIC EXCELLENCE.
- 

25 years of specialized service in  
international education.

---

Sponsored by the American  
Economic Association



For further information write:

The Director  
Economics Institute  
Campus Box 259  
University of Colorado  
Boulder, Colorado 80309  
(303) 492-7337  
Telex: 45-0385

*Please mention THE AMERICAN ECONOMIC REVIEW When Writing to Advertisers*



**MATHEMATICAL AND STATISTICAL PROGRAMMING PACKAGE FOR YOUR IBM PC**  
**FAST • EASY TO USE • POWERFUL**

# GAUSS™

**YOU'VE NEVER SEEN ANYTHING LIKE IT!**

**GAUSS** is a sophisticated mathematical and statistical programming package for the IBM PC and compatibles. It combines speed, power, and ease of use in one amazing program.

**GAUSS** allows you to do essentially anything you can do with a mainframe statistical package — and a lot more.

Personal computers are friendly, convenient, and inexpensive. So is **GAUSS**. **GAUSS** is not just a stripped-down mainframe program. **GAUSS** has been designed from the ground up to take advantage of all of the conveniences of a personal computer. After trying **GAUSS**, you may never use a mainframe again.

**GAUSS** comes with programs written in its matrix programming language that allow you to do most econometric procedures, including OLS, 2SLS, 3SLS, PROBIT, LOGIT, MAXIMUM LIKELIHOOD, and NON-LINEAR LEAST SQUARES.

In the current version, **GAUSS** will accept up to 90 variables in a regression. There is no limit on the number of observations.

**GAUSS** will do a regression with 10 independent variables and 800 observations in under 4 seconds — and with 50 variables and 10,000 observations in under 18 minutes. It will compute the maximum likelihood estimates of a binary logit model, with 10 variables and 1,000 observations, in 1-2 minutes, depending upon the number of iterations required.

**GAUSS** allows you to do complicated statistical procedures that you would never imagine trying on a mainframe. It is easy to program almost any routine, and **GAUSS** is so fast that it can do almost any job. But the nicest thing of all is that the cost of time on your personal computer is essentially zero!

**GAUSS** is an excellent teaching tool. It makes programming easy and allows students to focus on concepts and techniques.

If you can write it mathematically, you can write it in **GAUSS**. Furthermore, you can write it in **GAUSS** almost exactly the way you would write it mathematically.

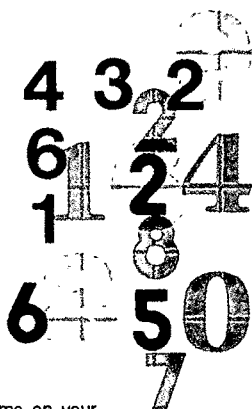
**GAUSS** is 10-15 times faster than other programs that use the 8087, and 15-100 times faster than other programs that do not use the 8087.

As in APL, single statements in **GAUSS** can accomplish what might take dozens of lines in another language. However, **GAUSS** provides you with additional powerful numerical operators and functions — especially for statistics and the solution of linear equations — that are not found in APL. And, of course, the syntax in **GAUSS** is much more natural (for most of us) than that in APL.

**GAUSS** has state-of-the-art numerical routines and random number generators.

**GAUSS** is extremely accurate. It allows you to do an entire regression in 19 digit accuracy. It will compute the Longley benchmark coefficients in 5 hundredths of a second with an average of 11 correct digits! (Try that on a mainframe!)

**GAUSS**, with its built-in random number generators and powerful functions and operators, is an excellent tool for doing simulations.



## **GAUSS and the 8087 NUMERIC DATA PROCESSOR GIVE YOU MINICOMPUTER PERFORMANCE ON YOUR DESKTOP.**

### **SPECIAL INTRODUCTORY OFFER**

With 30 Day Money

Back Guarantee ..... Reg. 395.00 **\$250.00**

**GAUSS** requires an IBM PC with at least 256K RAM, an 8087 NDP, 1 DS/DD disk drive, DOS 2.0 (or above).

IBM is trademark of IBM Corporation

Call or Write

### **APPLIED TECHNICAL SYSTEMS**

P.O. Box 6487, Kent, WA 98064  
(206) 631-6679

# OECD publications

ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT

## **National Accounts, Volume I: Main Aggregates, 1960-1983**

After a summary of the definitions of these main aggregates, this volume is divided into five parts. Part one contains graphs for each country showing growth in real terms of GDP and expenditure. Part two gives for the OECD as a whole and for each country the main aggregates in national currencies. Part three provides "growth triangles" for the main components of final expenditure. Part four gives a set of comparative tables (US dollars, volume and price indices, population and exchange rates), with totals for groups of countries and for the OECD as a whole. Finally, part five contains a set of comparative tables using PPPs.

1985 130 pages \$18.00

## **Geographical Distribution of Financial Flows to Developing Countries, 1980-1983**

The unique source of data on the sources, volume and terms of official aid flows to 110 developing countries. It also gives the origin, type, and inflow of other external financial resources as well as some basic economic information for each developing country covered.

1984 290 pages \$32.00

## **Costs and Margins in Banking: Statistical Supplement 1978-1982**

As a statistical supplement to the 1980 OECD report on *Costs and Margins in Banking: An International Survey*, this volume provides a unique tool for analyzing developments during the period 1978-1982.

1985 112 pages \$25.00

## **The Internationalisation of Banking: The Policy Issues**

*R. M. Pecchioli*

Examines the implications of the expansion of world credit and the emergence of banks as channels for balance-of-payments financing.

1983 222 pages \$22.00

## **Transfer Pricing and Multinational Enterprises: Three Taxation Issues**

The three reports published in this volume supplement the 1979 report of the same name. They consider problems for multinational enterprises resulting from the adjustment of transfer prices by tax authorities, transfer pricing in the banking sector, and the allocation of management and service costs for tax purposes.

1984 92 pages \$12.00

Available from:

**OECD Publications and Information Center**

1750-E Pennsylvania Avenue, N.W.

Washington, D.C. 20006-4582 Tel.: (202) 724-1857



# IRWIN

## Textbooks in Economics for 1985

### NEW TITLES

#### **Public Finance**

Harvey S. Rosen, Princeton University  
*Instructor's Manual*

#### **Antitrust Economics**

Roger D. Blair, University of Florida  
and David L. Kaserman, University  
of Tennessee

### REVISIONS

#### **Managerial Economics: Applied Microeconomics for Decision Making, 2nd Edition**

S. Charles Maurice and Charles W.  
Smithson, both of Texas A & M  
University  
*Workbook/Study Guide, Instructor's  
Manual w. Text Bank*

#### **Macroeconomics:**

#### **Microeconomics:**

#### **Analysis and Policy, 5th Edition**

Lloyd G. Reynolds, Yale University  
*paperbound w. Workbook/Study  
Guide, Instructor's Manual, and  
Manual of Tests for each volume*

#### **The New World of Economics: Explorations into the Human Experience, 4th Edition**

Gordon Tullock, George Mason Univer-  
sity, and Richard B. McKenzie, Clemson  
University  
*paperbound*

#### **Public Policies toward Business, 7th Edition**

William G. Shepherd, University of  
Michigan

#### **The Practice of Collective Bargaining, 7th Edition**

James P. Begin, Rutgers University, and  
Edwin F. Beal, Emeritus, University of  
Oregon  
*Instructor's Manual*

#### **Readings in Labor Eco- nomics and Labor Rela- tions, 5th Edition**

Richard L. Rowan (editor), The Wharton  
School, University of Pennsylvania  
*paperbound*

#### **Comparative Economic Systems: Models and Cases, 5th Edition**

Morris Bornstein, University of Michigan

Examination copies for adoption consideration available on request.  
Please indicate course title and text presently used.

Richard D. Irwin, Inc. Homewood, Illinois 60430

27 AUG 1985